Pre-training Models

Xu Tan
Microsoft Research Asia
xuta@microsoft.com

Outline

- Overview of pre-training
- Taxonomy of pre-training: context vs contrast
- More discussion about pre-training
- Summary

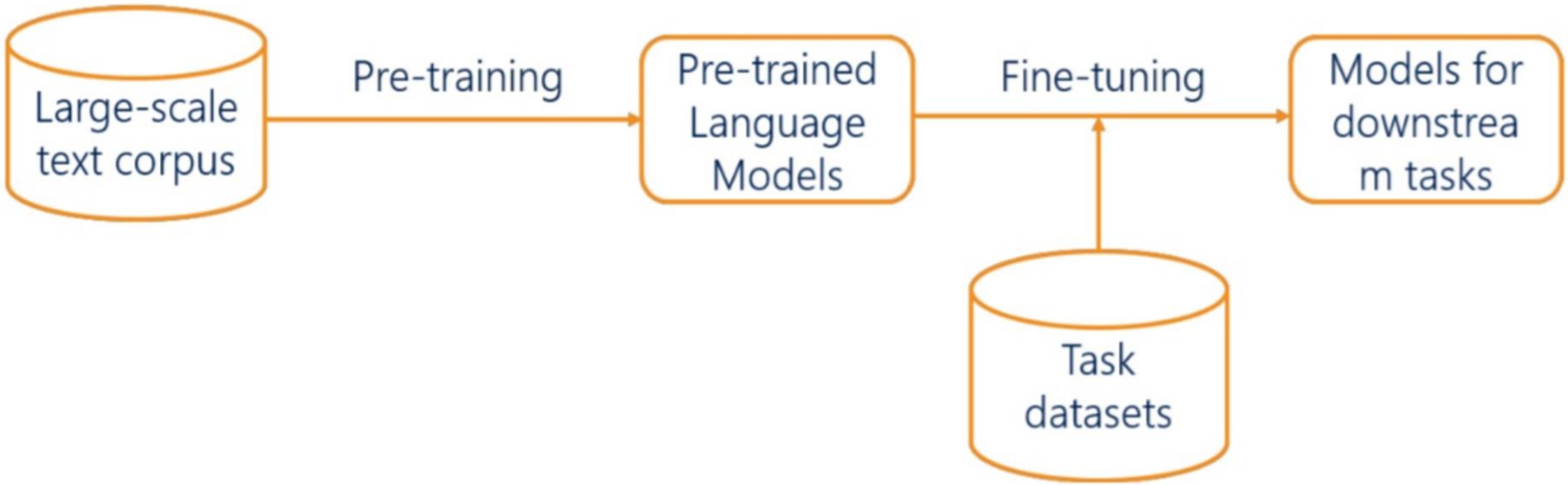
Pre-training

- What is pre-training?
  - The training in advance of standard training
- Why pre-training?
  - The standard training (data/model size) is not enough
- What to learn in pre-training?
  - Representation learning: more general, self-supervised
  - Task learning: more task specific, supervised
- When and where to apply pre-training?
  - Any tasks that data/model size are not enough
  - NLP, CV, Speech, and more

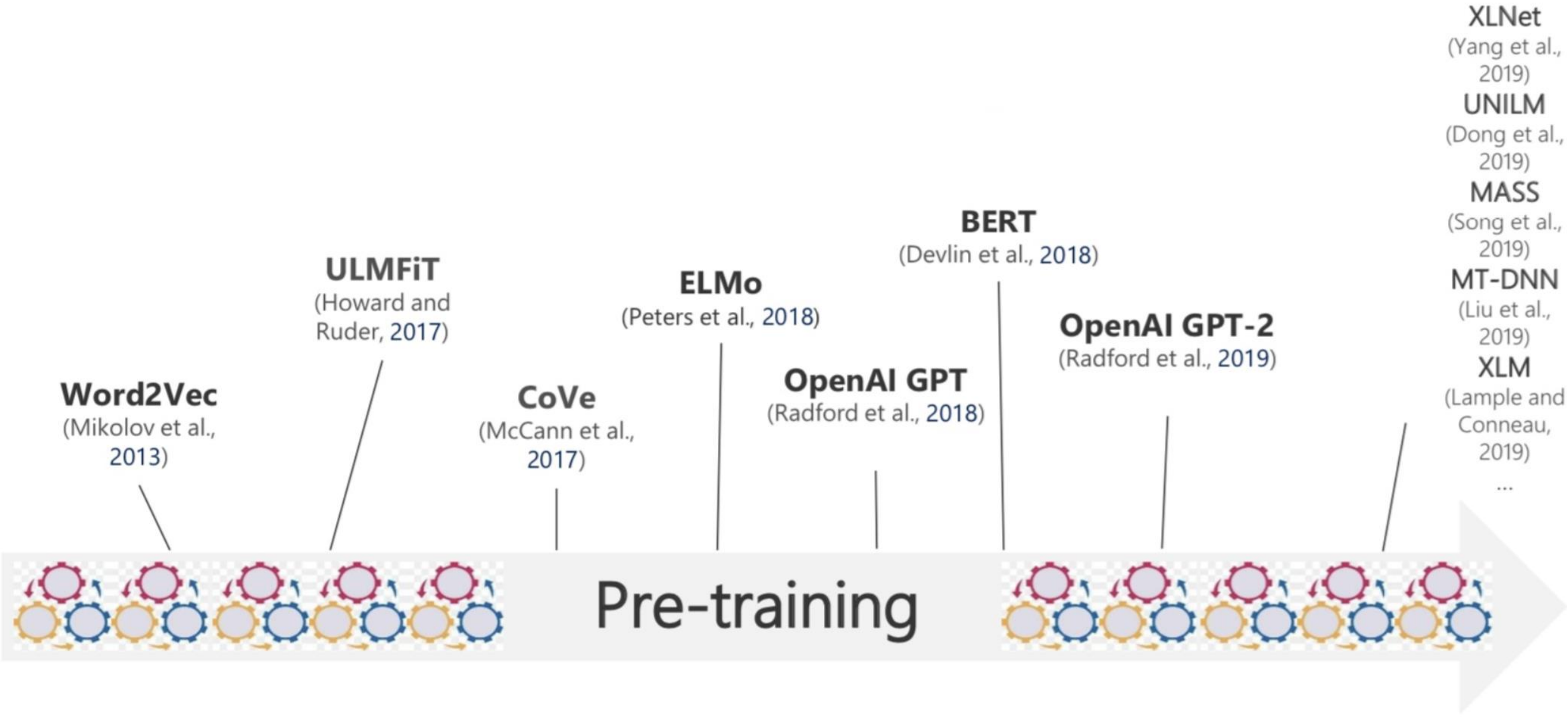
- **How to learn in pre-training?**

Pre-training in NLP

- Pre-training + Fine-tuning, a new paradigm of NLP



Pre-training in NLP



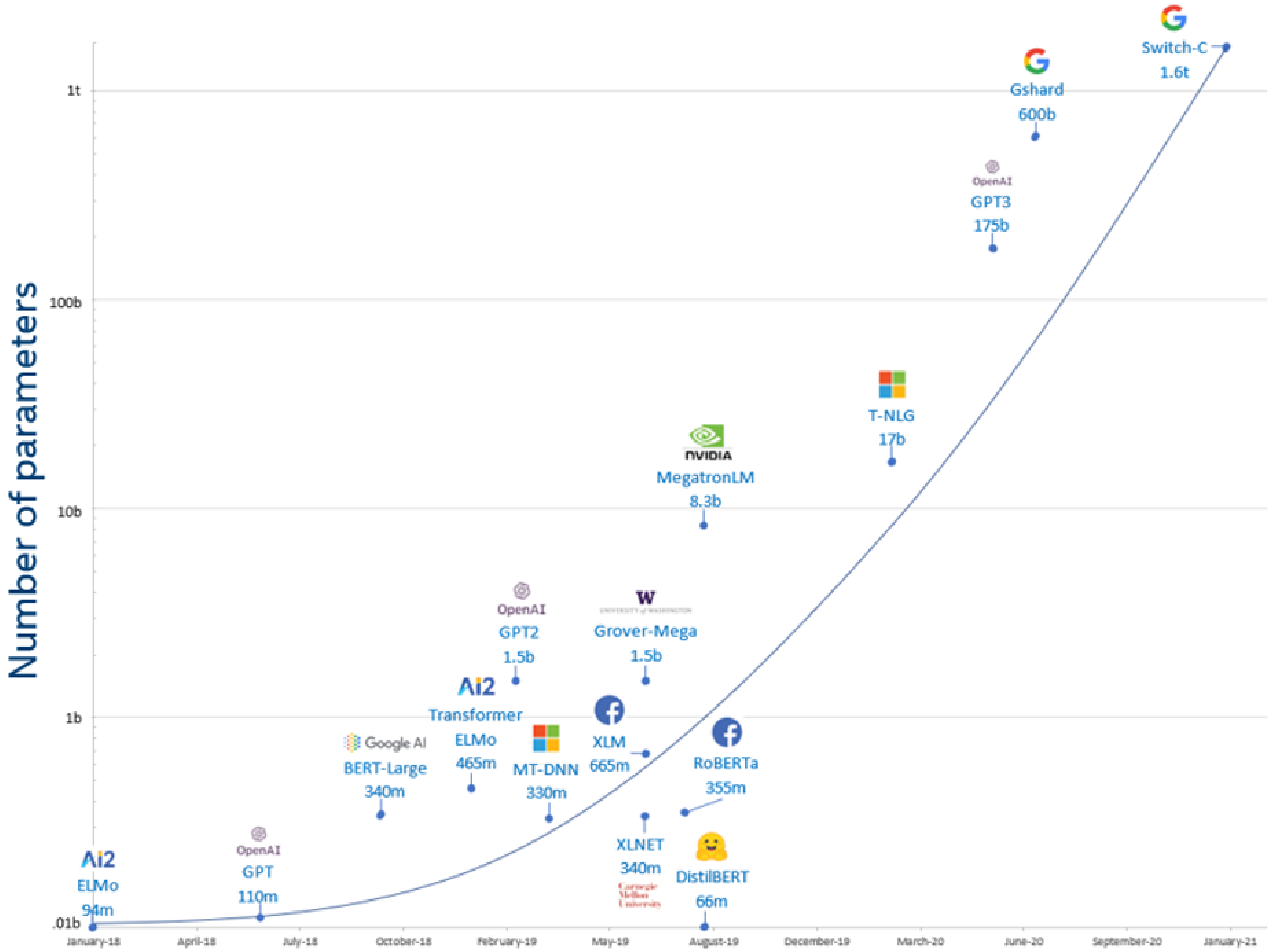
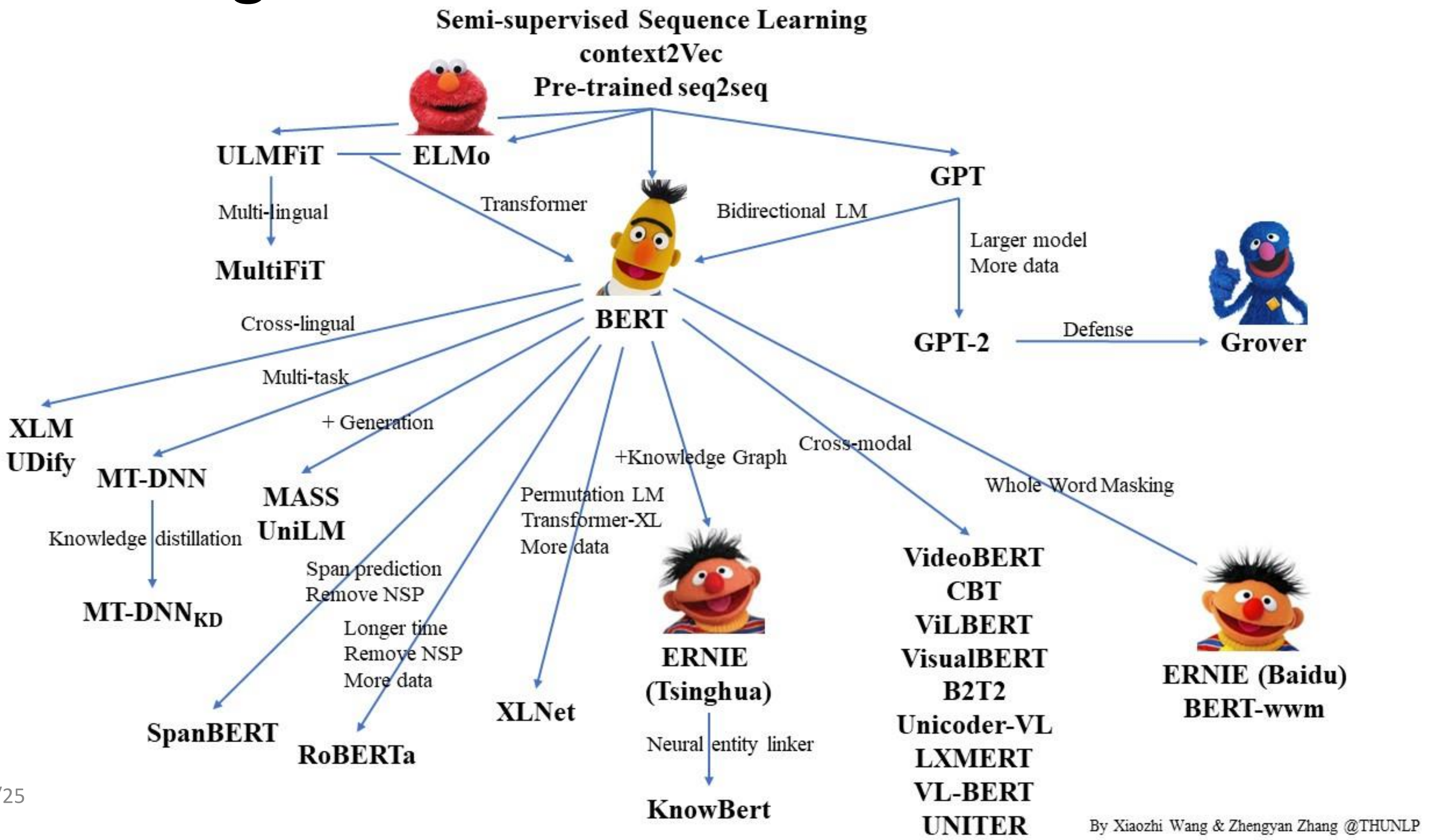
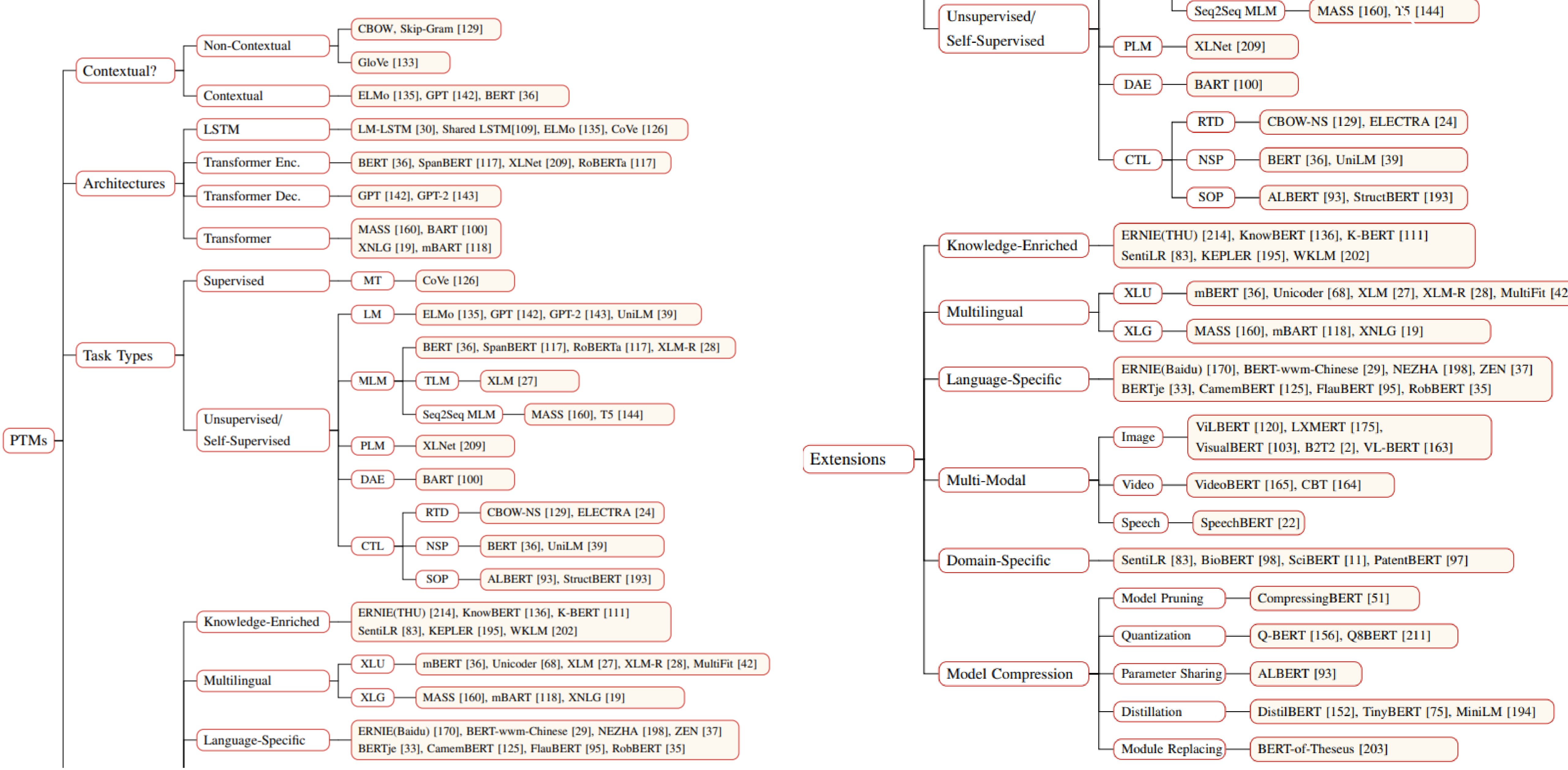


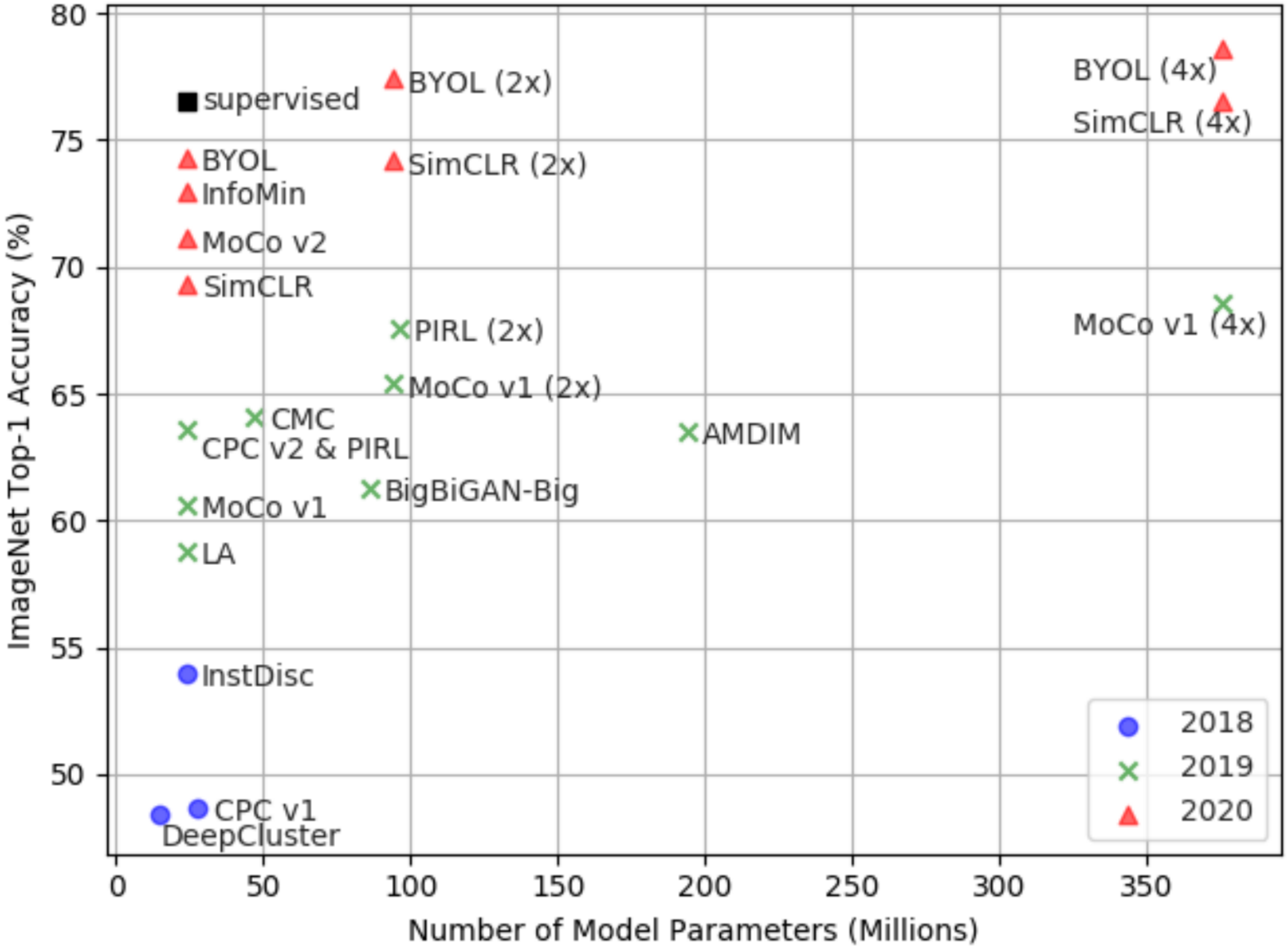
Figure 1: Exponential growth of number of parameters in DL models

Pre-training in NLP




PTMs
Contextual?
Non-Contextual
CBOW, Skip-Gram [129]
GloVe [133]
Contextual
ELMo [135], GPT [142], BERT [36]
Architectures
LSTM
LM-LSTM [30], Shared LSTM[109], ELMo [135], CoVe [126]
Transformer Enc.
BERT [36], SpanBERT [117], XLNet [209], RoBERTa [117]
Transformer Dec.
GPT [142], GPT-2 [143]
Transformer
MASS [160], BART [100]
XNLG [19], mBART [118]
Task Types
Supervised
MT
CoVe [126]
Unsupervised/
Self-Supervised
LM
ELMo [135], GPT [142], GPT-2 [143], UniLM [39]
MLM
BERT [36], SpanBERT [117], RoBERTa [117], XLM-R [28]
TLM
XLM [27]
Seq2Seq MLM
MASS [160], T5 [144]
PLM
XLNet [209]
DAE
BART [100]
CTL
RTD
CBOW-NS [129], ELECTRA [24]
NSP
BERT [36], UniLM [39]
SOP
ALBERT [93], StructBERT [193]
Knowledge-Enriched
ERNIE(THU) [214], KnowBERT [136], K-BERT [111]
SentiLR [83], KEPLER [195], WKLM [202]
Multilingual
XLU
mBERT [36], Unicoder [68], XLM [27], XLM-R [28], MultiFit [42]
XLG
MASS [160], mBART [118], XNLG [19]
Language-Specific
ERNIE(Baidu) [170], BERT-wwm-Chinese [29], NEZHA [198], ZEN [37]
BERTje [33], CamemBERT [125], FlauBERT [95], RobBERT [35]
Unsupervised/
Self-Supervised
Seq2Seq MLM
MASS [160], T5 [144]
PLM
XLNet [209]
DAE
BART [100]
CTL
RTD
CBOW-NS [129], ELECTRA [24]
NSP
BERT [36], UniLM [39]
SOP
ALBERT [93], StructBERT [193]
Extensions
Knowledge-Enriched
ERNIE(THU) [214], KnowBERT [136], K-BERT [111]
SentiLR [83], KEPLER [195], WKLM [202]
Multilingual
XLU
mBERT [36], Unicoder [68], XLM [27], XLM-R [28], MultiFit [42
XLG
MASS [160], mBART [118], XNLG [19]
Language-Specific
ERNIE(Baidu) [170], BERT-wwm-Chinese [29], NEZHA [198], ZEN [37]
BERTje [33], CamemBERT [125], FlauBERT [95], RobBERT [35]
Multi-Modal
Image
ViLBERT [120], LXMERT [175],
VisualBERT [103], B2T2 [2], VL-BERT [163]
Video
VideoBERT [165], CBT [164]
Speech
SpeechBERT [22]
Domain-Specific
SentiLR [83], BioBERT [98], SciBERT [11], PatentBERT [97]
Model Compression
Model Pruning
CompressingBERT [51]
Quantization
Q-BERT [156], Q8BERT [211]
Parameter Sharing
ALBERT [93]
Distillation
DistilBERT [152], TinyBERT [75], MiniLM [194]
Module Replacing
BERT-of-Theseus [203]

Pre-training in CV——Self-supervised



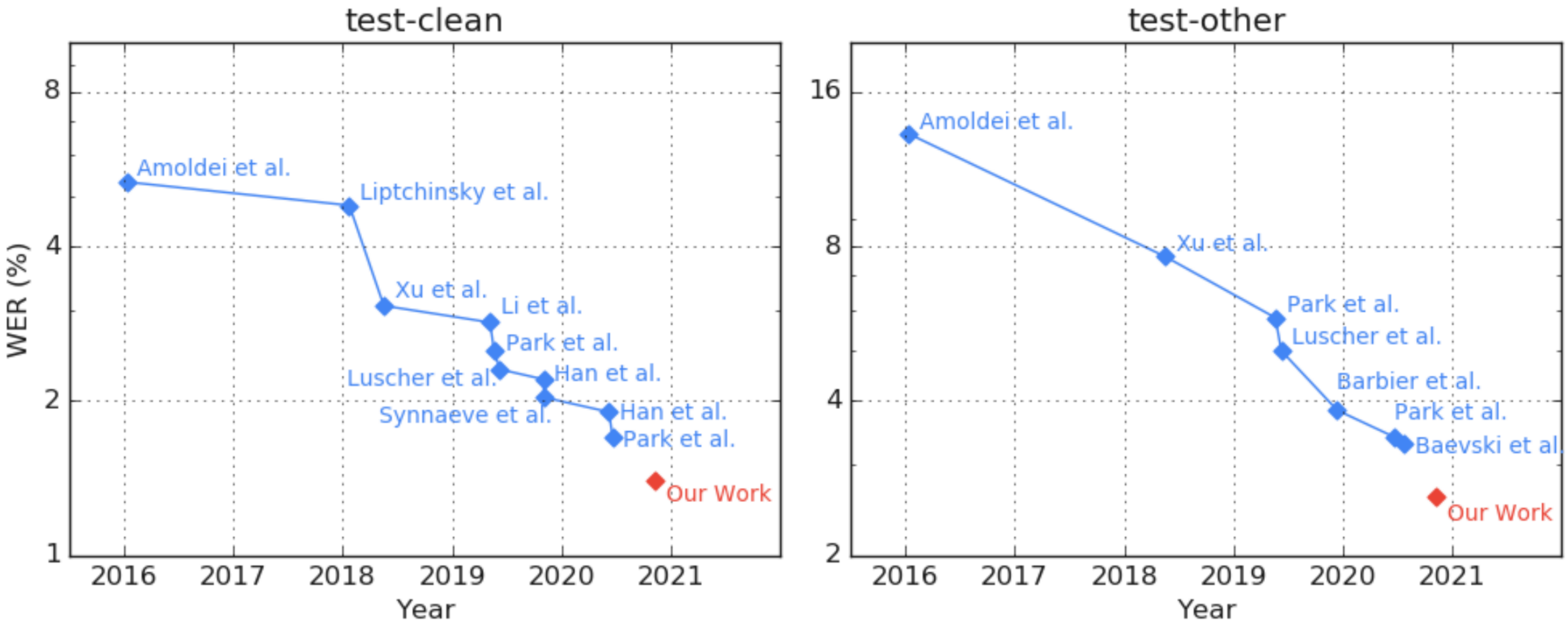
Top 1 accuracy on ImageNet with self-supervised pre-training

Pre-training in CV——Supervised/Semi-supervised

Rank	Model	Top 1 Accuracy	Top 5 Accuracy	Number of params	Extra Training Data	Code	Result	Year
1	**Meta Pseudo Labels** (EfficientNet-L2) Meta Pseudo Labels	90.2%	98.8%	480M	✓			2021
2	**Meta Pseudo Labels** (EfficientNet-B6-Wide) Meta Pseudo Labels	90%	98.7%	390M	✓			2021
3	**NFNet-F4+** High-Performance Large-Scale Image Recognition Without Normalization	89.2%		527M	✓			2021
4	**ALIGN** (EfficientNet-L2) Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision	88.64%	98.67%	480M	✓			2021
5	**EfficientNet-L2-475** (SAM) Sharpness-Aware Minimization for Efficiently Improving Generalization	88.61%		480M	✓			2020
6	**ViT-H/14** An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale	88.55%		632M	✓			2020
7	**FixEfficientNet-L2** Fixing the train-test resolution discrepancy: FixEfficientNet	88.5%	98.7%	480M	✓			2020
8	**NoisyStudent** (EfficientNet-L2) Self-training with Noisy Student improves ImageNet classification	88.4%	98.7%	480M	✓			2020
9	**Mixer-H/14** (JFT-300M pre-train) MLP-Mixer: An all-MLP Architecture for Vision	87.94%			✓			2021
10	**ViT-L/16** An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale	87.76%		307M	✓			2020

Top 1 accuracy on ImageNet with supervised/semi-supervised pre-training

Pre-training in Speech



SOTA WER on LibriSpeech with self-supervised and semi-supervised training

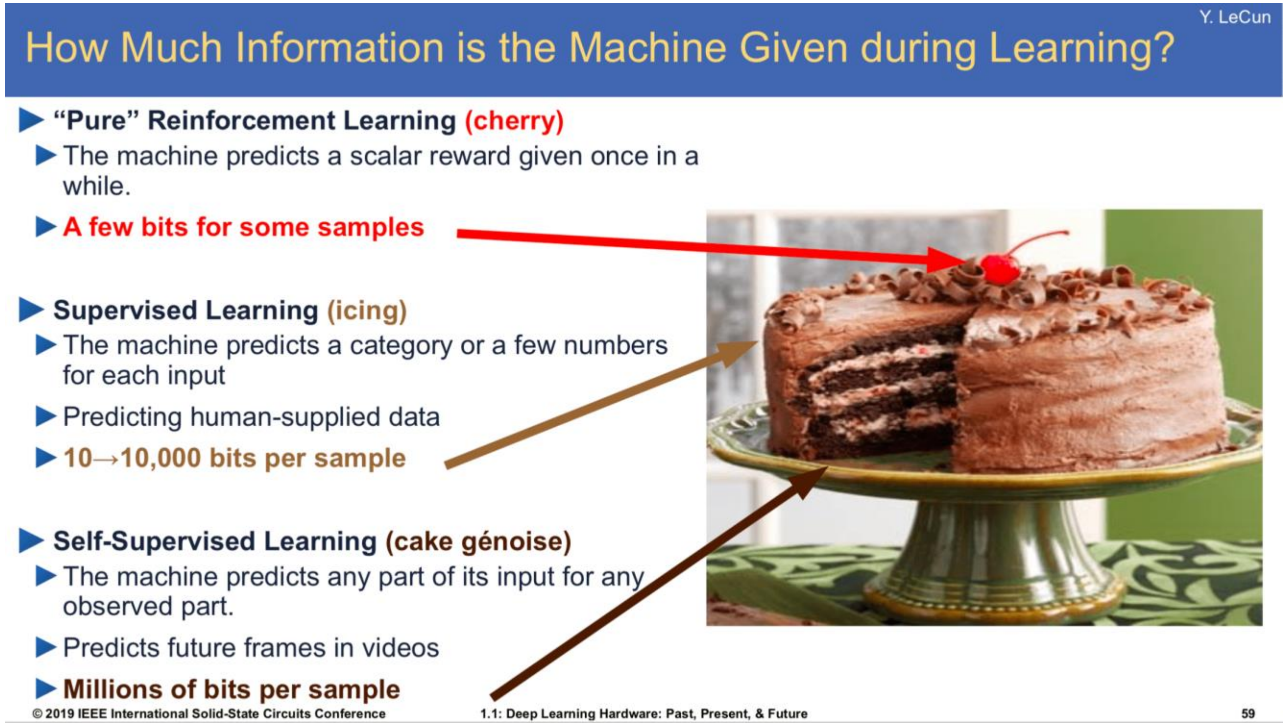
Pre-training in Speech

Rank	Model	Word Error Rate (WER)	Extra Training Data	Code	Result	Year
1	**Conformer + Wav2vec 2.0 + SpecAugment-based Noisy Student Training with Libri-Light** Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition	1.4	✓			2020
2	**Conv + Transformer + wav2vec2.0 + pseudo labeling** Self-training and Pre-training are Complementary for Speech Recognition	1.5	✓			2020
3	**ContextNet + SpecAugment-based Noisy Student Training with Libri-Light** Improved Noisy Student Training for Automatic Speech Recognition	1.7	✓			2020
4	**SpeechStew** (1B) SpeechStew: Simply Mix All Available Speech Recognition Data to Train One Large Neural Network	1.7	✕			2021
5	**Multistream CNN with Self-Attentive SRU** ASAPP-ASR: Multistream CNN and Self-Attentive SRU for SOTA Speech Recognition	1.75	✕			2020
6	**wav2vec 2.0 with Libri-Light** wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations	1.8	✓			2020
7	**ContextNet** (L) ContextNet: Improving Convolutional Neural Networks for Automatic Speech Recognition with Global Context	1.9	✕			2020
8	**Conformer** (L) Conformer: Convolution-augmented Transformer for Speech Recognition	1.9	✕			2020

SOTA WER on LibriSpeech with self-supervised and semi-supervised training

How to learn in pre-training

- Learning paradigm
  - Supervised learning
  - Unsupervised learning
  - Semi-supervised learning
  - Reinforcement learning
  - Transfer learning
  - Self-supervised learning
- Pre-training
  - In this talk, we focus more on self-supervised learning
  - Context based and contrast based

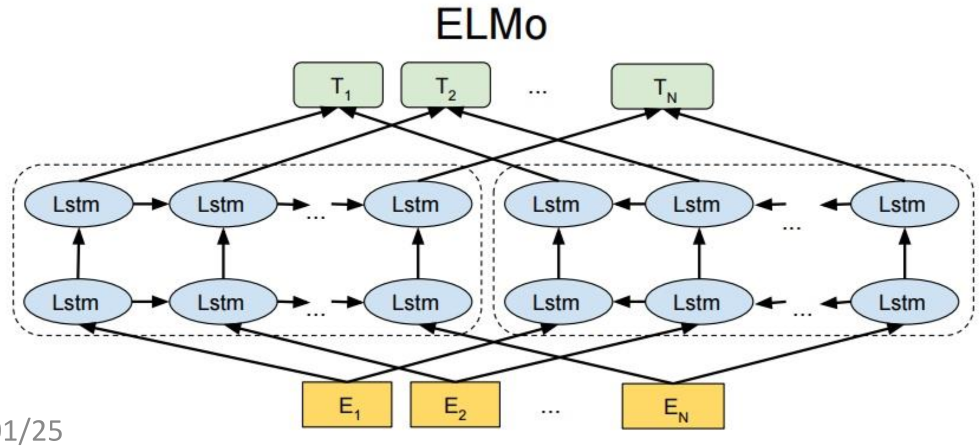


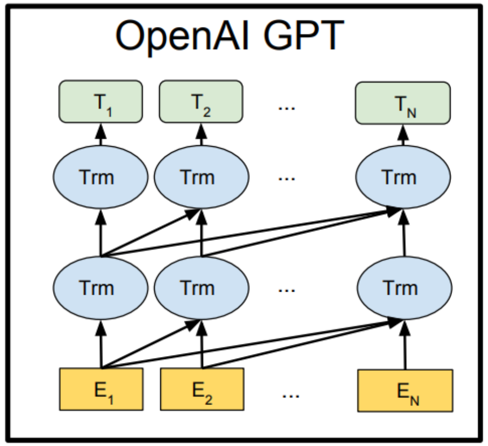
Context based vs Contrast based

- Context based
  - Autoregressive Language Model (LM): ELMo [3], GPT-1/2/3 [4,5,6]
  - Denoising Auto-Encoder (DAE): MLM (BERT[7], RoBERTa[9], ERNIE[21,23], UniLM[14], XLM [15]), Seq2SeqMLM (MASS [11], T5 [17], ProphetNet [43], BART[12])
  - Permuted Language Model (PLM): XLNet [10], MPNet [27]
- Contrast based
  - Context-Instance Contrast
    - Predict Relative Position (PRP): Jigsaw, Rotation Angle [45], Sentence Order Prediction (ALBERT [19], StructBERT [20])
    - Maximize Mutual Information (MI): Deep InfoMax/InforWord [28], AMDIM [29], Contrastive Predictive Coding [30] (wav2vec [41,42]), Replaced Token Detection (word2vec [1], ELECTRA[18])
  - Context-Context Contrast
    - DeepCluster [32], CMC [31], MoCo [34,37], SimCLR [35,38], BYOL [36], Next Sentence Prediction (BERT [7])

Context based: LM

- Language model $\mathcal{L}_{\mathrm{LM}} = -\sum_{t=1}^{T} \log p(x_t|\mathbf{x}_{<t})$
  - Natural, Joint probability estimation
  - Left to right, no bidirectional context information
- ELMo [3], GPT [4], GPT-2 [5], GPT-3 [6]
  - ELMo: NAACL 2018 best paper, GPT-3: NeurIPS 2020 best paper
  - GPT-3 has 175 billion parameters, the largest model before (1.7 Trillion, Switch Transformer [46])





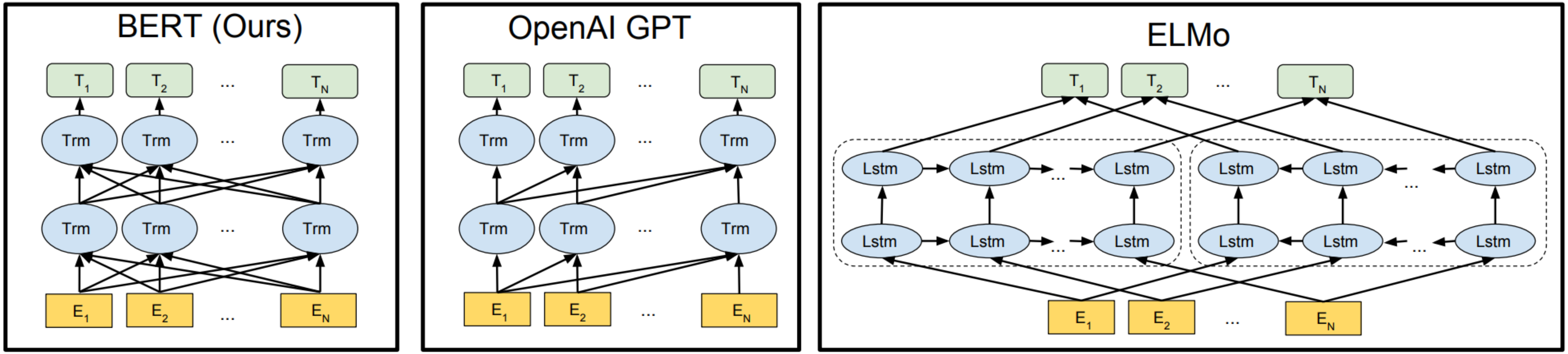
Context based: DAE

- Denoising Auto-Encoder
  - DAE: Noisy Input, reconstruct whole clean input $$\mathcal{L}_{\text{DAE}} = -\sum_{t=1}^{T} \log p(x_t|\hat{\mathbf{x}}, \mathbf{x}_{<t})$$
  - MLM: Noisy Input (with mask tokens), reconstruct mask tokens $$\mathcal{L}_{\text{MLM}} = -\sum_{\hat{x}\in m(\mathbf{x})} \log p\left(\hat{x}|\mathbf{x}_{\backslash m(\mathbf{x})}\right)$$
  - Seq2SeqMLM: Noisy Input (with mask tokens), reconstruct mask tokens, with encoder-decoder framework $$\mathcal{L}_{\text{S2SMLM}} = -\sum_{t=i}^{j} \log p\left(x_t|\mathbf{x}_{\backslash \mathbf{x}_{i:j}}, \mathbf{x}_{i:t-1}\right)$$

- BERT [7], MASS [11], RoBERTa [9], XLM [15], ERNIE [21,23], UniLM [14], ProphetNet [43], T5 [17], BART [12]

Context based: MLM——BERT

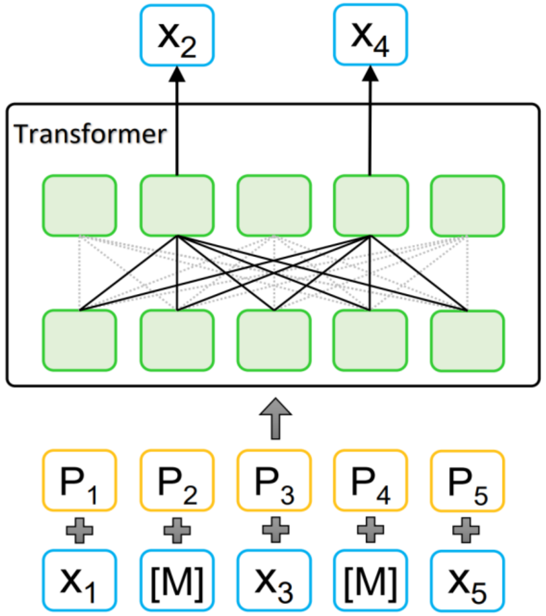
- BERT [7]: Bidirectional transformer, vs GPT [4,5,6], ELMo [3]

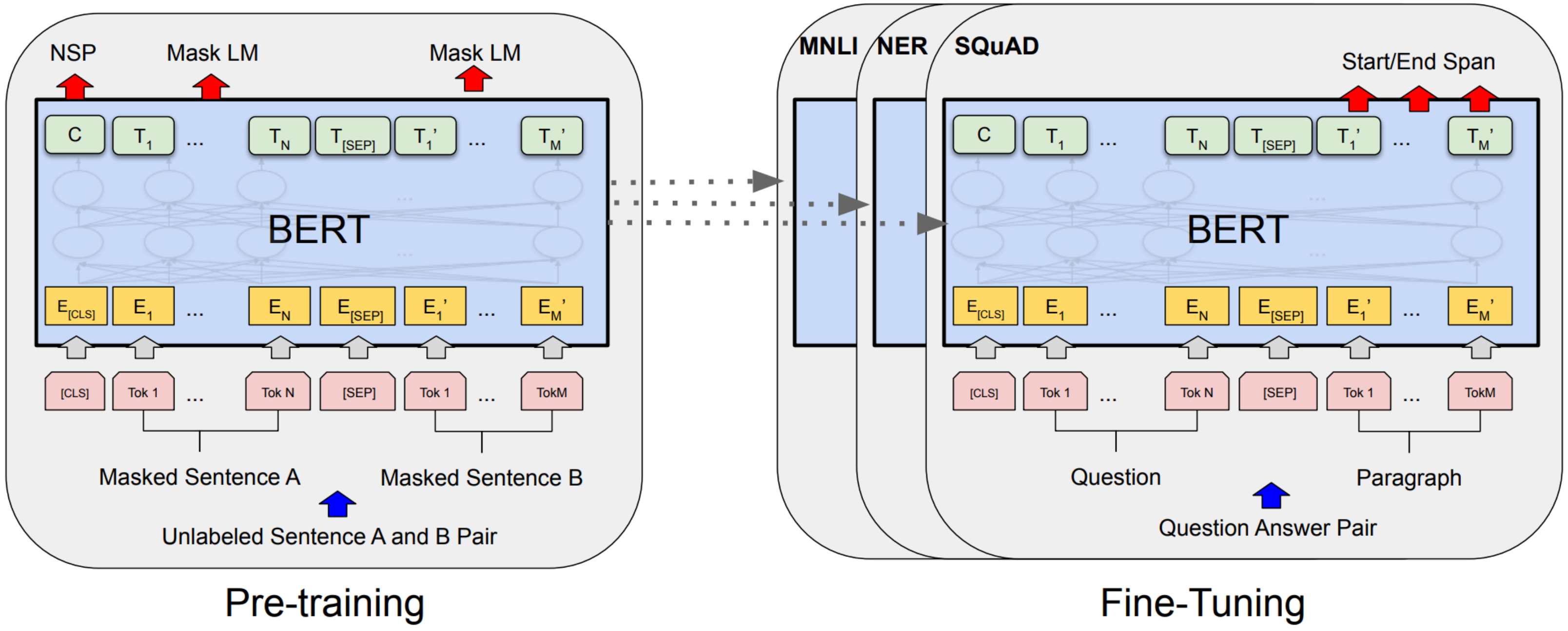


[7] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *NAACL 2019*

Context based: MLM——BERT

- BERT [7]: Bidirectional transformer

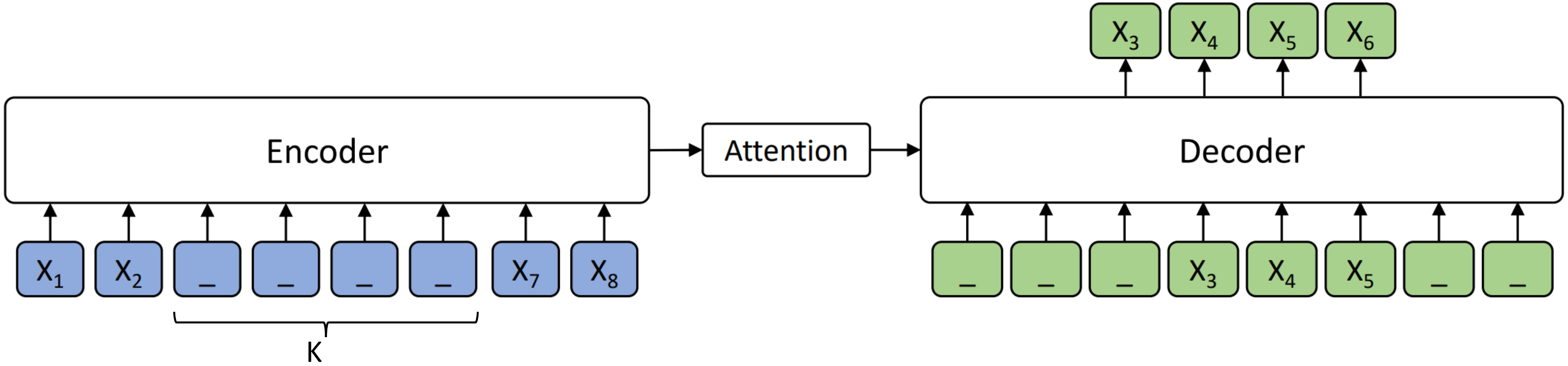




[7] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *NAACL 2019*

Context based: Seq2SeqMLM——MASS

- MASS: MAsked Sequence to Sequence pre-training [11]
  - MASS is carefully designed to jointly pre-train the encoder and decoder



- Mask k consecutive tokens (a sentence segment)
  - **Force the decoder to attend on the source representations, i.e., encoder-decoder attention.**
  - Force the encoder to extract meaningful information from the sentence.
  - Develop the decoder with the ability of language modeling.

[11] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu. MASS: Masked sequence to sequence pre-training for language generation. ICML 2019

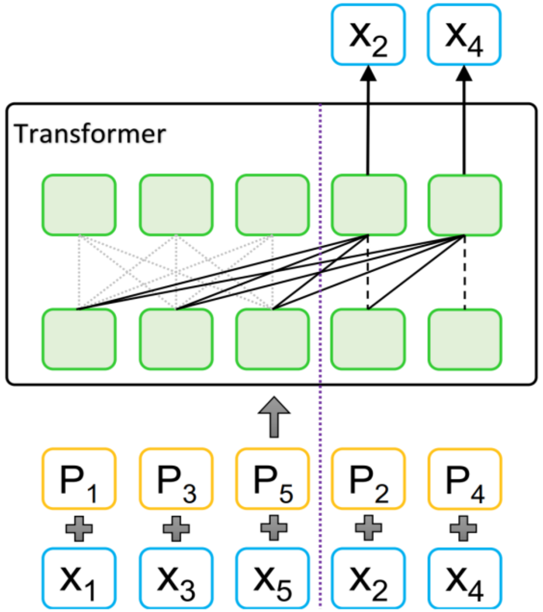
Context based: PLM

- Permuted Language Model $\mathcal{L}_{\text{PLM}} = -\sum_{t=1}^{T} \log p(z_t | \mathbf{z}_{<t})$
    - A generalized autoregressive language model, random permute the sentence order, better use language model for pre-training
    - Combine the advantages of LM and MLM
        - LM: only left context, MLM: bidirectional context
        - LM: conditional dependent, MLM: conditional independent
- XLNet [10], MPNet [27]

[10] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In NeurIPS, pages 5754–5764, 2019.
[27] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, Tie-Yan Liu. MPNet: Masked and permuted pre-training for language understanding. NeurIPS 2020.

Context based: PLM——XLNet

- Key designs in XLNet
  - Autoregressive model, use permuted language model (PLM) to introduce bidirectional context
  - Two-stream self-attention to decide the position of next predicted token
  - Use Transformer-XL to incorporate long context

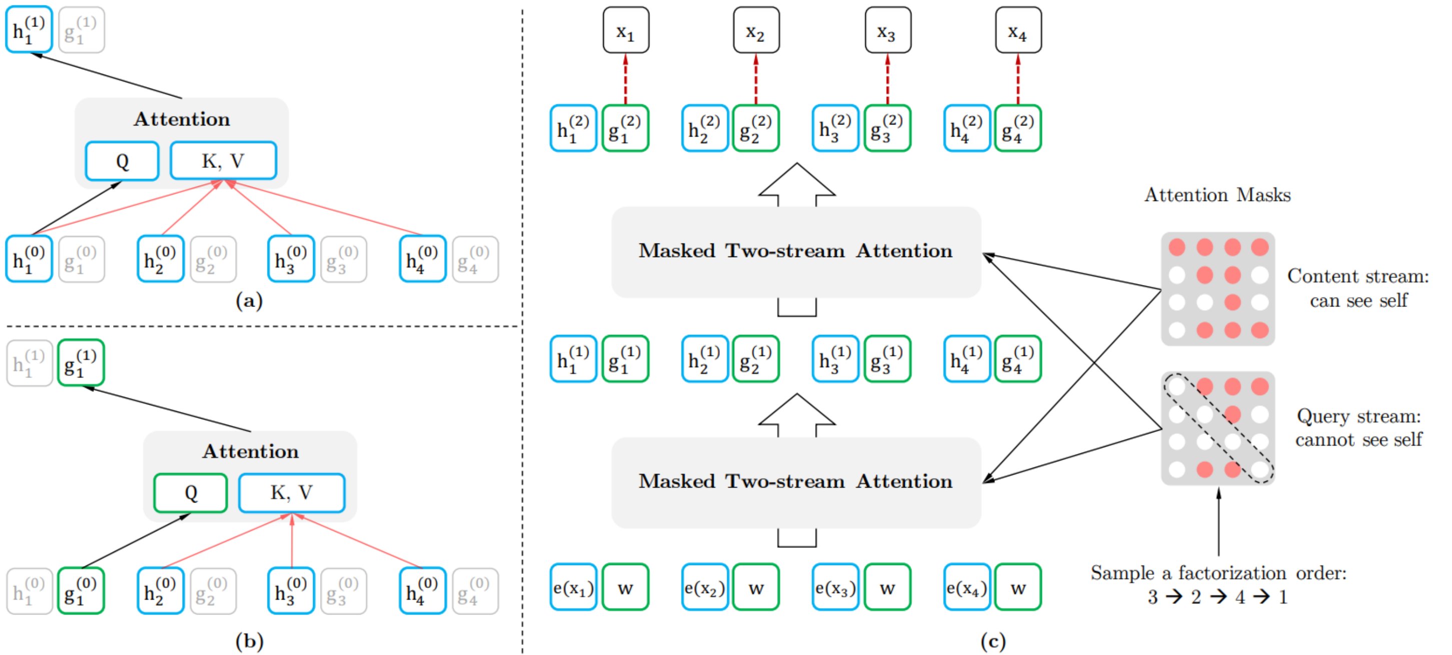


$$\log P(x;\theta) = \mathbb{E}_{z\in\mathcal{Z}_n} \sum_{t=c+1}^{n} \log P(x_{z_t} | x_{z_{<t}}; \theta)$$

Context based: PLM——XLNet

- Two-stream self-attention
  - Content stream: build content hidden, same as GPT/BERT in Transformer
  - Query stream: token prediction, use position as input to decide which token to predict

$$g_{z_t}^{(m)} \leftarrow \text{Attention}(\mathrm{Q} = g_{z_t}^{(m-1)}, \mathrm{KV} = \mathbf{h}_{\mathbf{z}_{<t}}^{(m-1)}; \theta), \quad (\text{query stream: use } z_t \text{ but cannot see } x_{z_t})$$
$$h_{z_t}^{(m)} \leftarrow \text{Attention}(\mathrm{Q} = h_{z_t}^{(m-1)}, \mathrm{KV} = \mathbf{h}_{\mathbf{z}_{\leq t}}^{(m-1)}; \theta), \quad (\text{content stream: use both } z_t \text{ and } x_{z_t}).$$



Context based: PLM——XLNet

- Transformer-XL
  - Recurrence mechanism: cache and reuse the representation of previous segment

$$h_{z_t}^{(m)} \leftarrow \text{Attention}(\mathrm{Q} = h_{z_t}^{(m-1)}, \mathrm{KV} = \left[\tilde{\mathbf{h}}^{(m-1)}, \mathbf{h}_{\mathbf{z}_{\leq t}}^{(m-1)}\right]; \theta)$$

  - Relative position embedding
    - Do not care the absolute position, but only relative position

$$\begin{aligned}\mathbf{A}_{i,j}^{\text{abs}} &= \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{U}_j}_{(b)} \\ &+ \underbrace{\mathbf{U}_i^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{E}_{x_j}}_{(c)} + \underbrace{\mathbf{U}_i^\top \mathbf{W}_q^\top \mathbf{W}_k \mathbf{U}_j}_{(d)}.\end{aligned}$$

$$\begin{aligned}\mathbf{A}_{i,j}^{\text{rel}} &= \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,E} \mathbf{E}_{x_j}}_{(a)} + \underbrace{\mathbf{E}_{x_i}^\top \mathbf{W}_q^\top \mathbf{W}_{k,R} \mathbf{R}_{i-j}}_{(b)} \\ &+ \underbrace{u^\top \mathbf{W}_{k,E} \mathbf{E}_{x_j}}_{(c)} + \underbrace{v^\top \mathbf{W}_{k,R} \mathbf{R}_{i-j}}_{(d)}.\end{aligned}$$

Context based: PLM——XLNet

- The advantage of XLNet (the task is sentence classification)
  - Bidirectional context, vs GPT

Objective	Modeling
LM (GPT)	log P(is | the task)
PLM (XLNet)	log P(is| the task) + log P(is | sentence classification)

  - Dependency between predicted tokens, vs BERT

Objective	Modeling
MLM (BERT)	log P(sentence| the task is) + log P(classification| the task is)
PLM (XLNet)	log P(sentence| the task is) + log P(classification| the task is sentence)

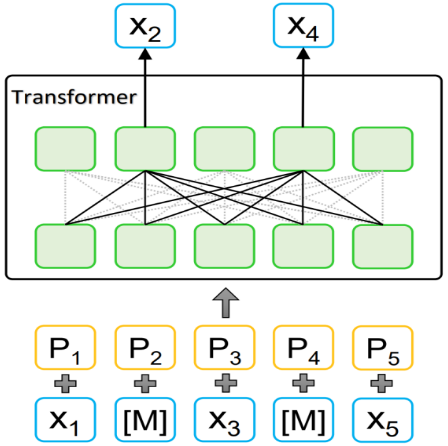
Context based: MLM+PLM——MPNet

- The pros and cons of BERT and XLNet
  - “the task is sentence classification”, predict token “sentence” and “classification”

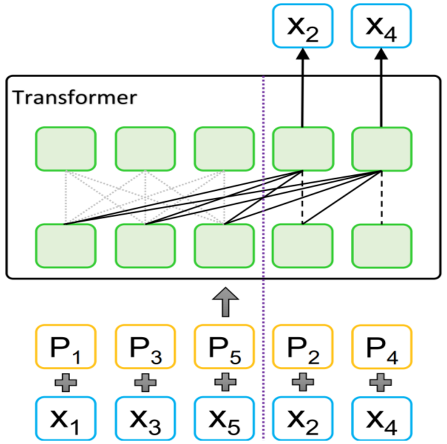
Objective	Modeling
MLM (BERT)	log P(sentence | the task is [M] [M]) + log P(classification | the task is [M] [M])
PLM (XLNet)	log P(sentence | the task is) + log P(classification | the task is sentence)

full position information

dependency



MLM (BERT)



PLM (XLNet)

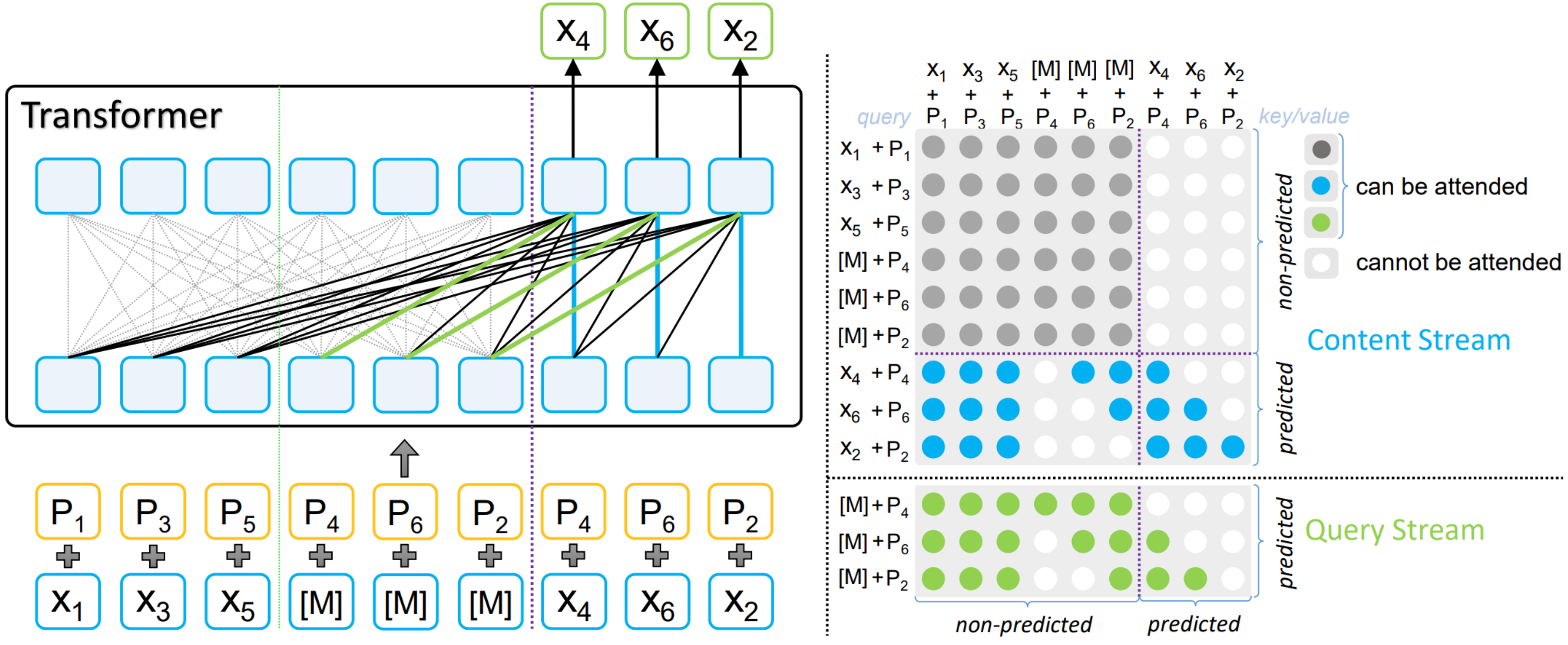
- Two aspects
  - Output dependency: dependency among the masked/predicted tokens
  - Input consistency: position information between pre-training and fine-tuning

	Output Dependency	Input Consistency
MLM (BERT)	×	√
PLM (XLNet)	√	×

MPNet: inherit their advantages but avoid their limitations

Context based: MLM+PLM——MPNet

- **Autoregressive prediction** (avoid the limitation in BERT)
  - Each predicted token condition on previous predicted tokens to ensure **output dependency**
- **Position compensation** (avoid the limitation in XLNet)
  - Each predicted token can see full position information to ensure **input consistency**



(a) (b)

Context based: MLM+PLM——MPNet

- The advantages of MPNet

Objective	Modeling
MLM (BERT)	log P(sentence | the task is [M] [M]) + log P(classification | the task is [M] [M])
PLM (XLNet)	log P(sentence | the task is) + log P(classification | the task is sentence)
MPNet	log P(sentence | the task is [M] [M]) + log P(classification | the task is sentence [M])

- Position compensation, input consistency, **vs. PLM (XLNet)**
  - MPNet knows 2 tokens to predict, instead of 3 tokens like “sentence pair classification”
- Autoregressive prediction, output dependency, **vs. MLM (BERT)**
  - MPNet can better predict “classification” given previous token “sentence”, instead of predicting “answering” as if to predict “question answering”

Context based: MLM+PLM——MPNet

- The advantages of MPNet

Objective	Modeling
MLM (BERT)	log P(sentence | the task is [M] [M]) + log P(classification | the task is [M] [M])
PLM (XLNet)	log P(sentence | the task is) + log P(classification | the task is sentence)
MPNet	log P(sentence | the task is [M] [M]) + log P(classification | the task is sentence [M])

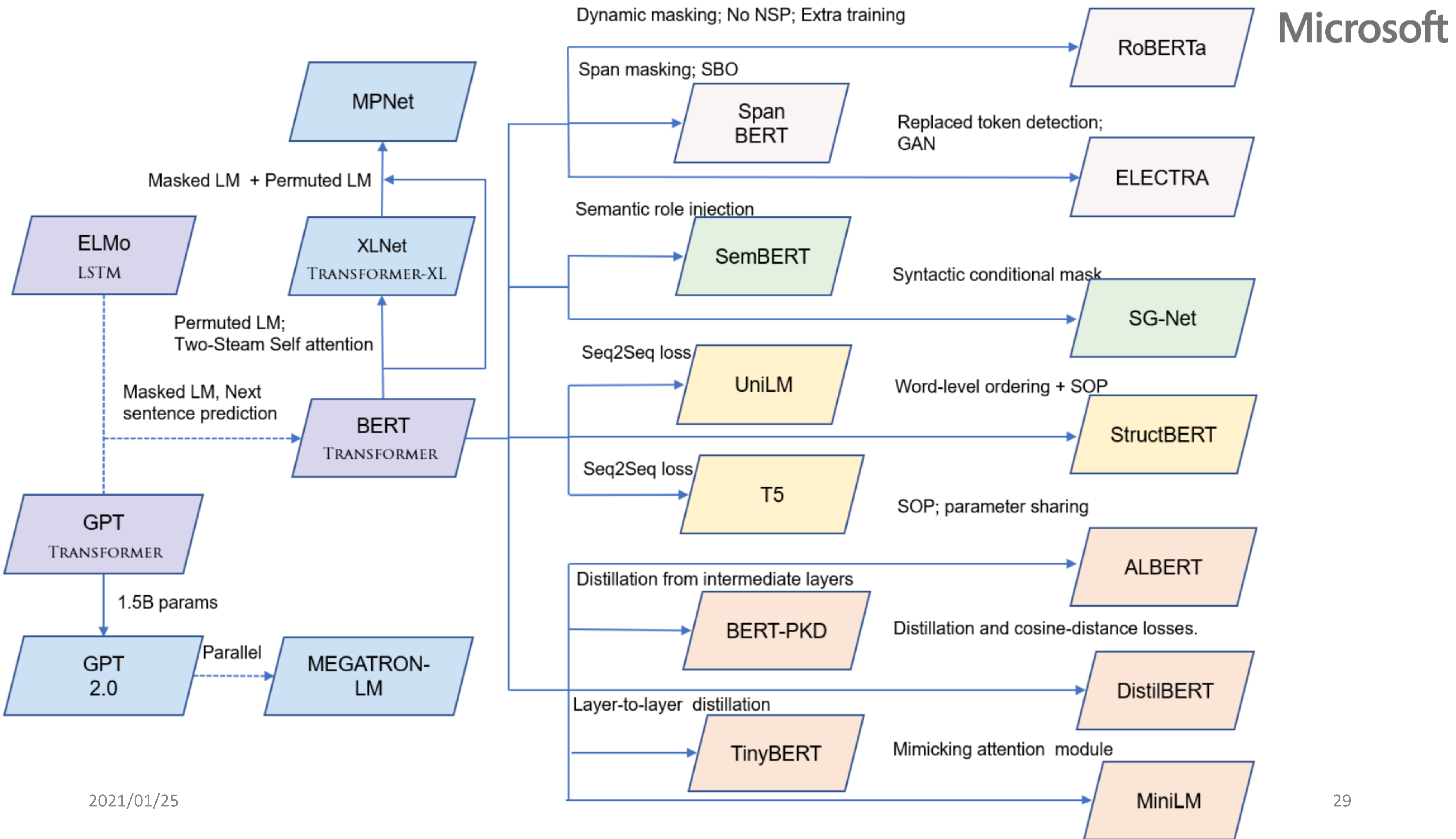
- How much conditional information is used on average to predict a masked token? (assume all objectives mask and predict 15% tokens)

Objective	Formulation	#Tokens	#Positions
MLM (BERT)	$\sum_{t=c+1}^{n} \log P(x_{z_t} \| x_{z_{<=c}}, M_{z_{>c}}; \theta)$	85%	100%
PLM (XLNet)	$\sum_{t=c+1}^{n} \log P(x_{z_t} \| x_{z_{<t}}; \theta)$	92.5%	92.5%
MPNet	$\sum_{t=c+1}^{n} \log P(x_{z_t} \| x_{z_{<t}}, M_{z_{>c}}; \theta)$	92.5%	100%

Inherit their advantages

Avoid their limitations

MPNet uses the most information to predict tokens


Microsoft
MPNet
Masked LM + Permuted LM
ELMo
LSTM
XLNet
TRANSFORMER-XL
Permuted LM;
Two-Steam Self attention
Masked LM, Next
sentence prediction
BERT
TRANSFORMER
GPT
TRANSFORMER
1.5B params
GPT
2.0
Parallel
MEGATRON-
LM
Dynamic masking; No NSP; Extra training
RoBERTa
Span masking; SBO
Span
BERT
Replaced token detection;
GAN
ELECTRA
Semantic role injection
SemBERT
Syntactic conditional mask
SG-Net
Seq2Seq loss
UniLM
Word-level ordering + SOP
StructBERT
Seq2Seq loss
T5
SOP; parameter sharing
ALBERT
Distillation from intermediate layers
BERT-PKD
Distillation and cosine-distance losses.
DistilBERT
Layer-to-layer distillation
TinyBERT
Mimicking attention module
MiniLM

Context based vs Contrast based

- Context based
  - Autoregressive Language Model (LM): ELMo [3], GPT-1/2/3 [4,5,6]
  - Denoising Auto-Encoder (DAE): MLM (BERT[7], RoBERTa[9], ERNIE[21,23], UniLM[14], XLM [15]), Seq2SeqMLM (MASS [11], T5 [17], ProphetNet [43], BART[12])
  - Permuted Language Model (PLM): XLNet [10], MPNet [27]
- Contrast based
  - Context-Instance Contrast
    - Predict Relative Position (PRP): Jigsaw, Rotation Angle [45], Sentence Order Prediction (ALBERT [19], StructBERT [20])
    - Maximize Mutual Information (MI): Deep InfoMax/InforWord [28], AMDIM [29], Contrastive Predictive Coding [30] (wav2vec [41,42]), Replaced Token Detection (word2vec [1], ELECTRA[18])
  - Context-Context Contrast
    - DeepCluster [32], CMC [31], MoCo [34,37], SimCLR [35,38], BYOL [36], Next Sentence Prediction (BERT [7])

Contrast based

- Basic idea: learn from contrast
  - Tell what is, and tell what is not

$$\mathcal{L}_N = -\mathbb{E}_{x,y+,y-}\left[\log \frac{\exp(s\left(x, y^{+}\right))}{\exp(s(x, y^{+})) + \sum_{j=1}^{N-1} \exp\left(s(x, y_j^{-})\right)}\right]$$

- Different contrast granularities
  - Context-Instance Contrast
  - Context-Context Contrast

Contrast based: Context-Instance Contrast

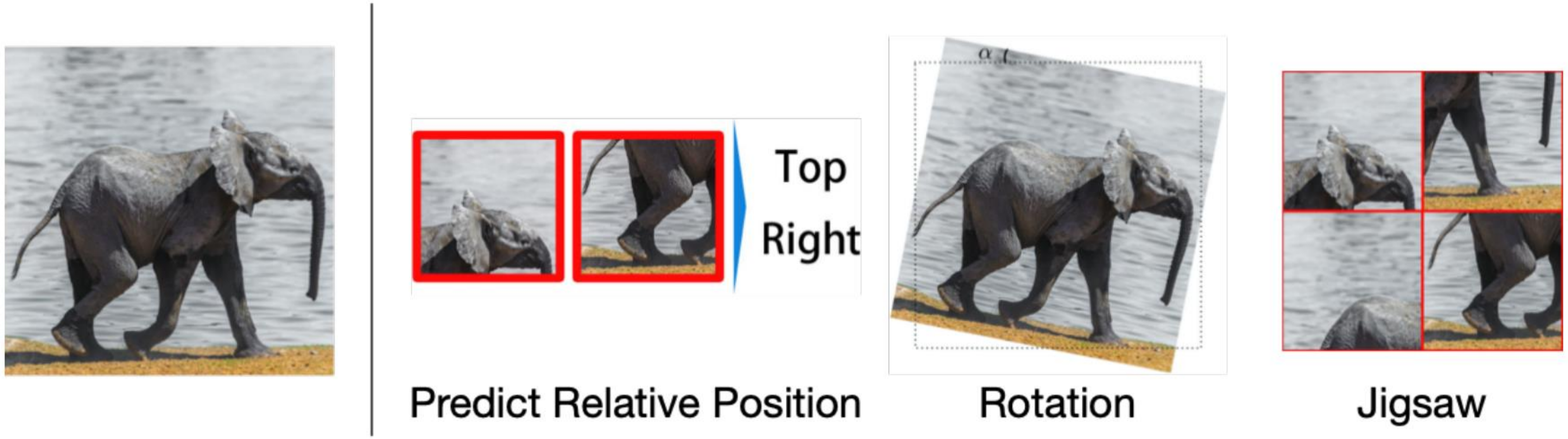
- Context-Instance Contrast: Global-local contrast: the local feature of a sample and its global context representation
  - Patches to their image, sentences to their paragraph, words to their sentence, and nodes to their neighborhoods.

$$\mathcal{L}_N = -\mathbb{E}_{x,y+,y-}\left[\log \frac{\exp(s\left(x, y^{+}\right))}{\exp(s(x, y^{+})) + \sum_{j=1}^{N-1} \exp\left(s(x, y_j^{-})\right)}\right]$$

  - Predict Relative Position (PRP): Jigsaw, Rotation Angle [45], Sentence Order Prediction (ALBERT [19], StructBERT [20])
  - Maximize Mutual Information (MI): Deep InfoMax/InforWord [28], AMDIM [29], Contrastive Predictive Coding [30] (wav2vec [41,42]), Replaced Token Detection (word2vec [1], ELECTRA [18])

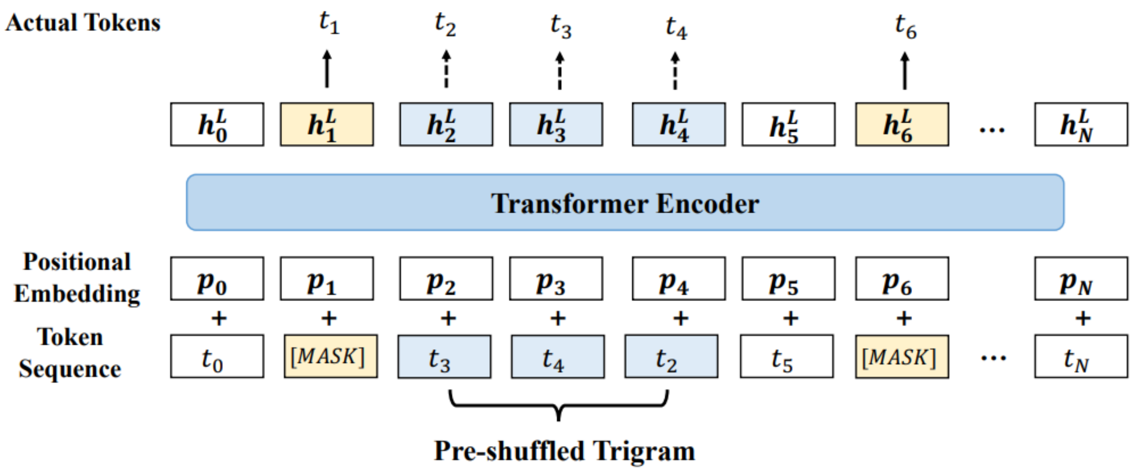
Contrast based: Context-Instance Contrast

- Predict Relative Position (PRP)
  - Jigsaw, rotation angle, relative position [45]

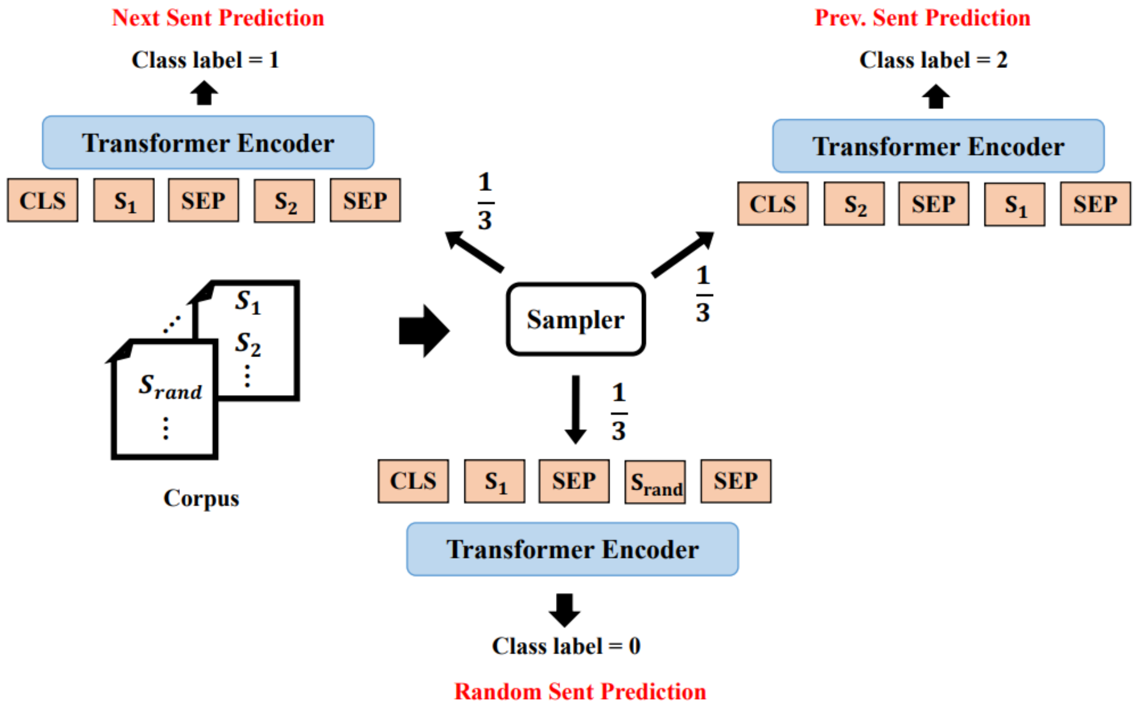


Contrast based: Context-Instance Contrast

- Predict Relative Position (PRP)
  - Next Sentence Prediction (BERT [7])
  - Sentence Order Prediction (ALBERT[19], StructBERT [20])



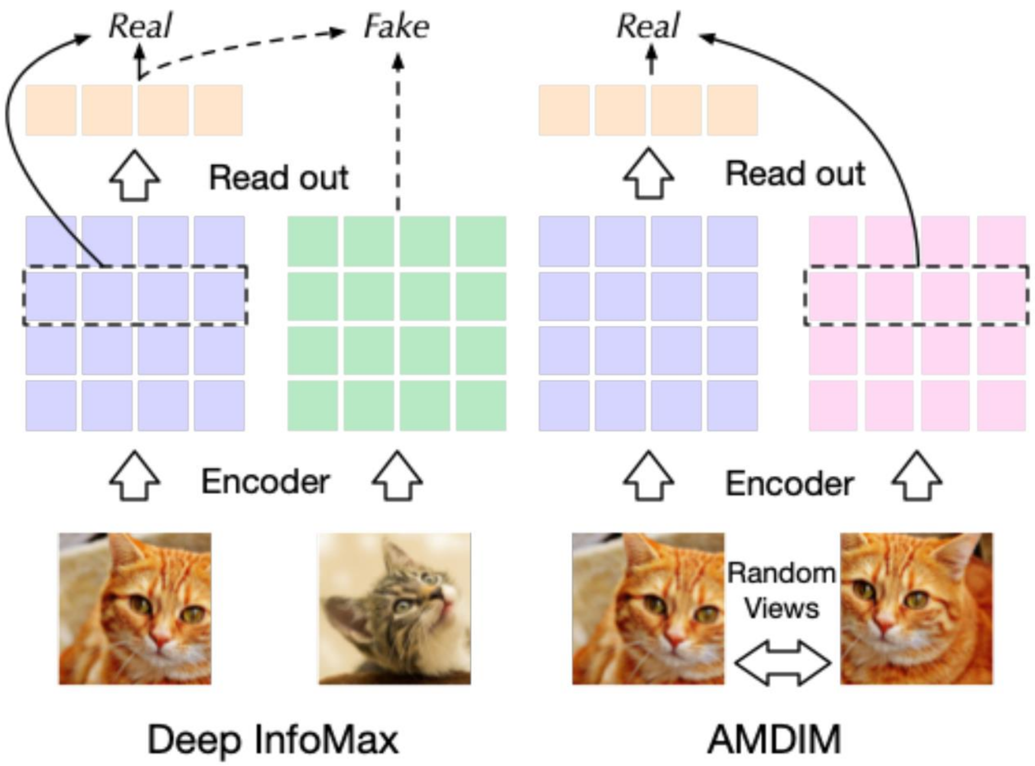
(a) Word Structural Objective



(b) Sentence Structural Objective

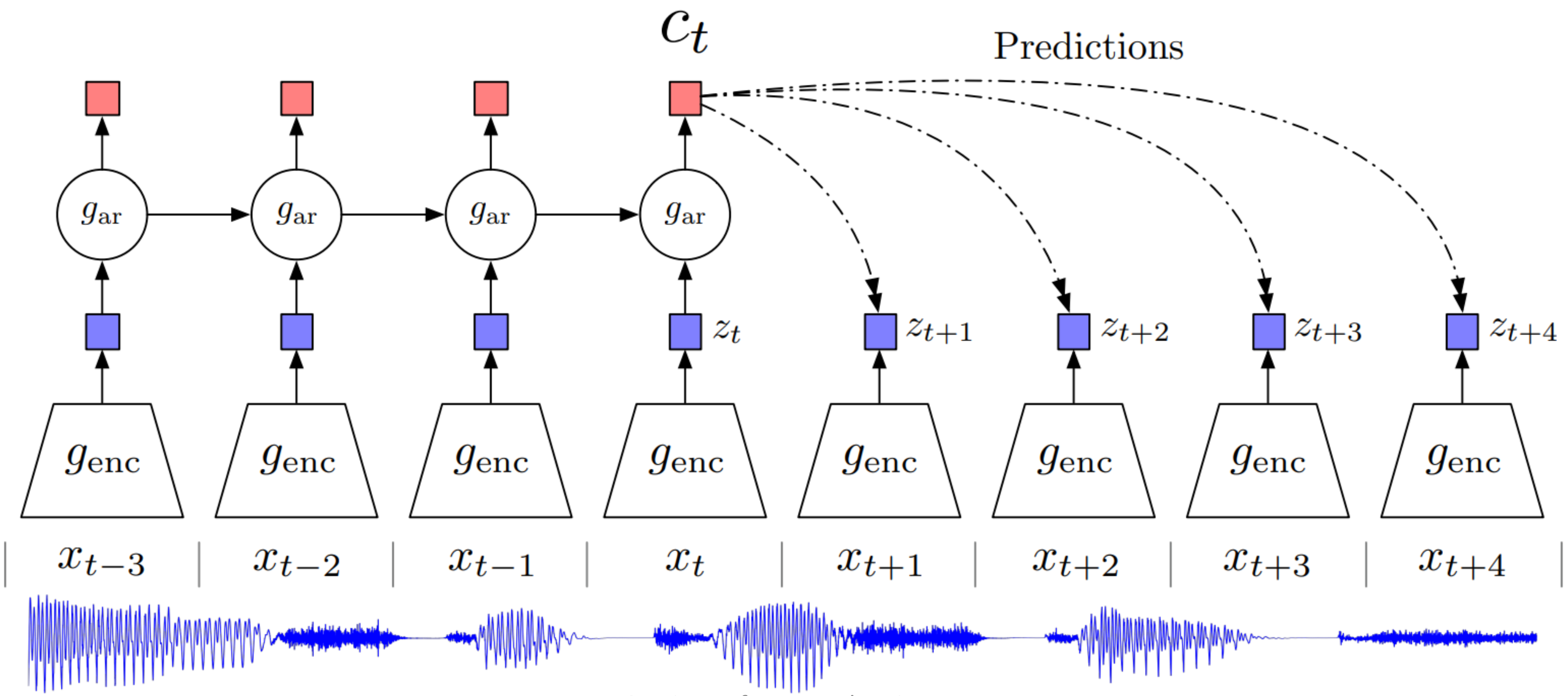
Contrast based: Context-Instance Contrast

- Maximize Mutual Information (MI)
  - Deep InfoMax/InfoWord [28], AMDIM [29]



Contrast based: Context-Instance Contrast

- Maximize Mutual Information (MI)
  - Contrastive Predictive Coding [30]

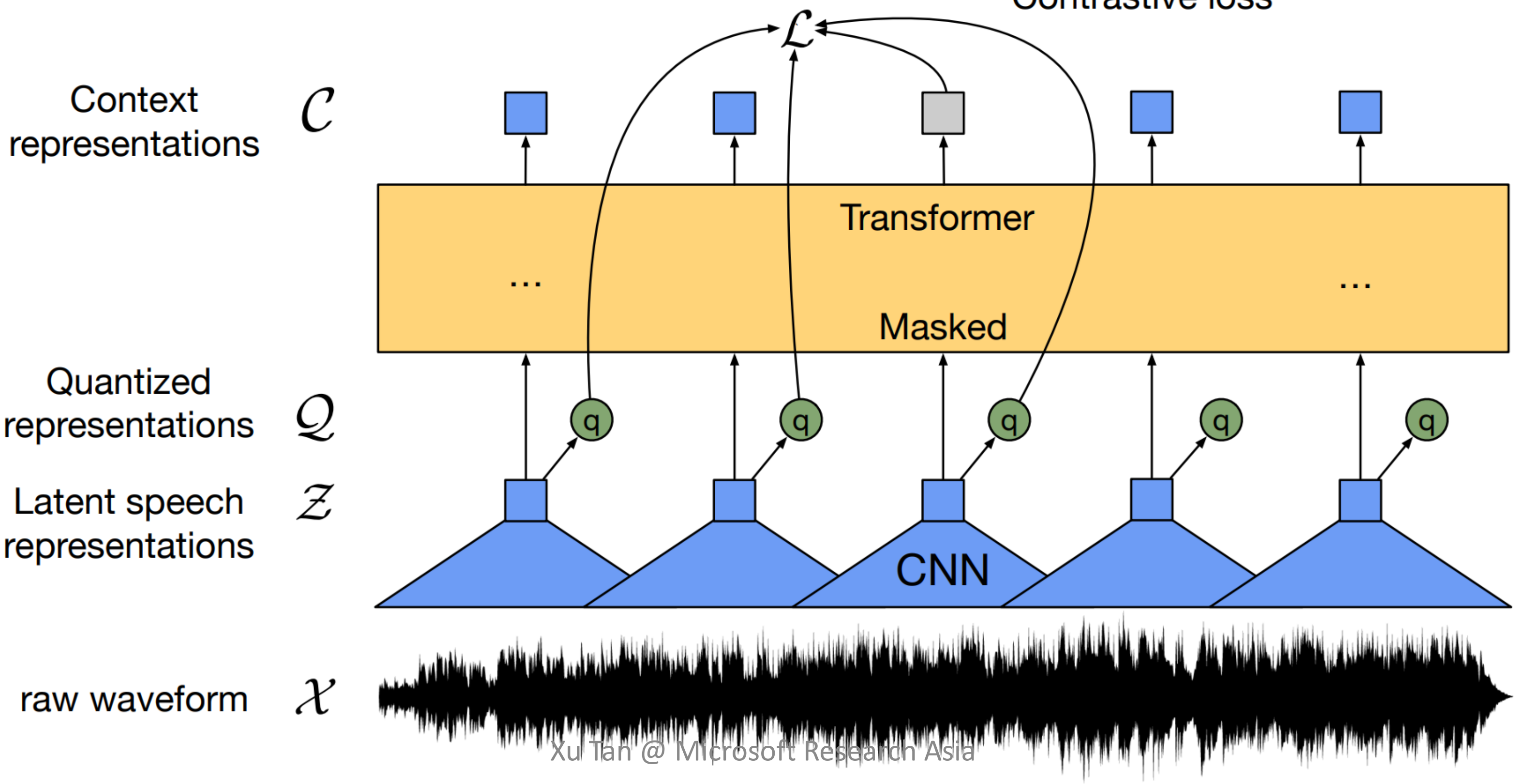


Contrast based: Context-Instance Contrast

- Maximize Mutual Information (MI)
  - Wav2vec /Wav2vec 2.0 [41,42]

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

Contrastive loss

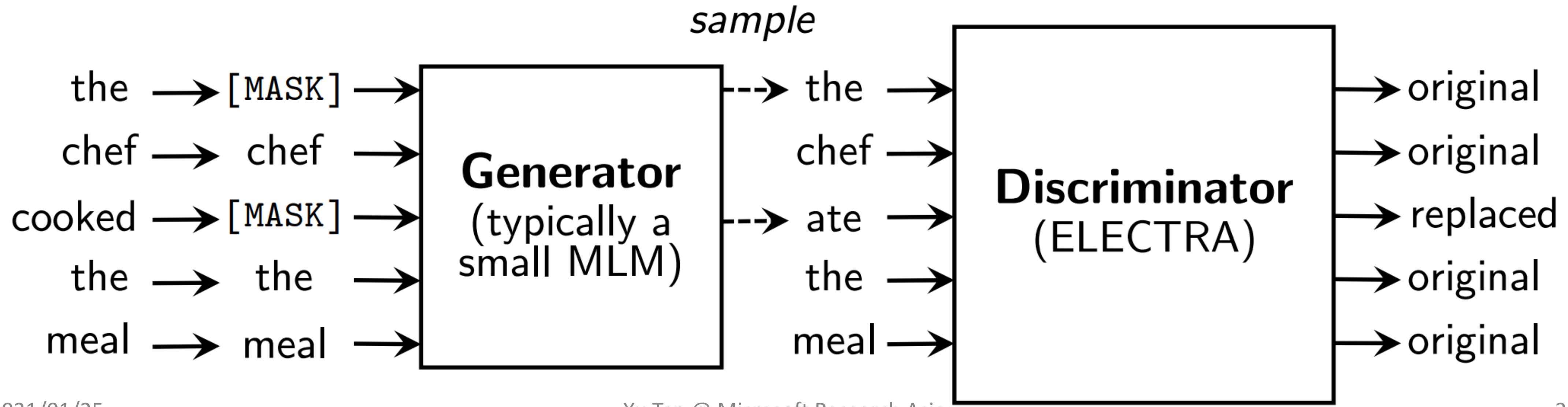


Contrast based: Context-Instance Contrast

- Maximize Mutual Information (MI)
  - Replaced Token Detection (word2vec [1], ELECTRA [18])

$$\mathcal{L}_{\text{RTD}} = -\sum_{t=1}^{T} \log p(y_t|\hat{\mathbf{x}})$$

$$\mathcal{L}_{\text{Disc}}(\boldsymbol{x}, \theta_D) = \mathbb{E}\left(\sum_{t=1}^{n} -\mathbb{1}(x_t^{\text{corrupt}} = x_t) \log D(\boldsymbol{x}^{\text{corrupt}}, t) - \mathbb{1}(x_t^{\text{corrupt}} \neq x_t) \log(1 - D(\boldsymbol{x}^{\text{corrupt}}, t))\right)$$

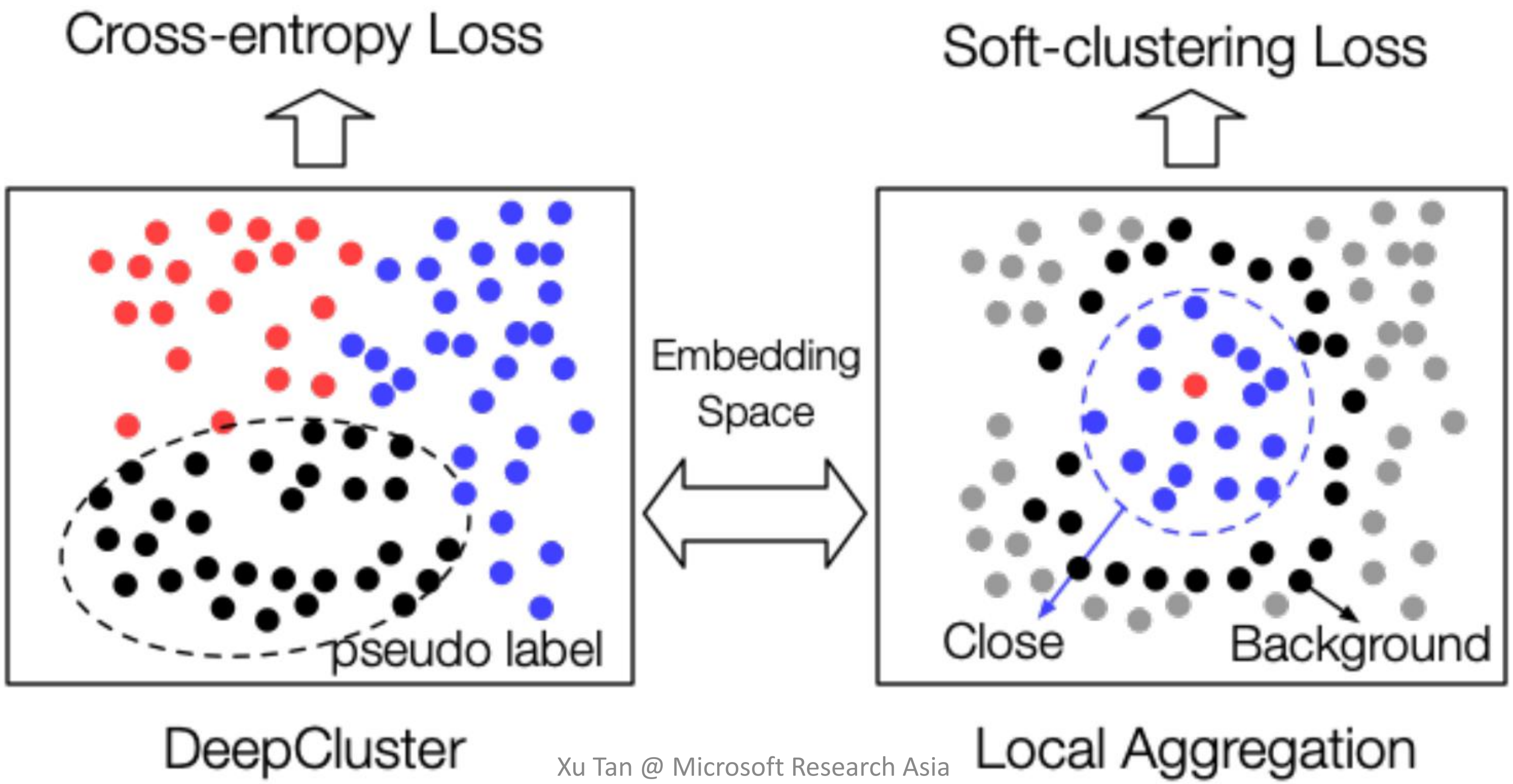


Contrast based: Context-Context Contrast

- Context-Context Contrast: the relationships between the global representations of different samples
  - Cluster-based Discrimination: DeepCluster [32]
  - Instance Discrimination: CMC [31], MoCo [34,37], SimCLR [35,38], BYOL [36]

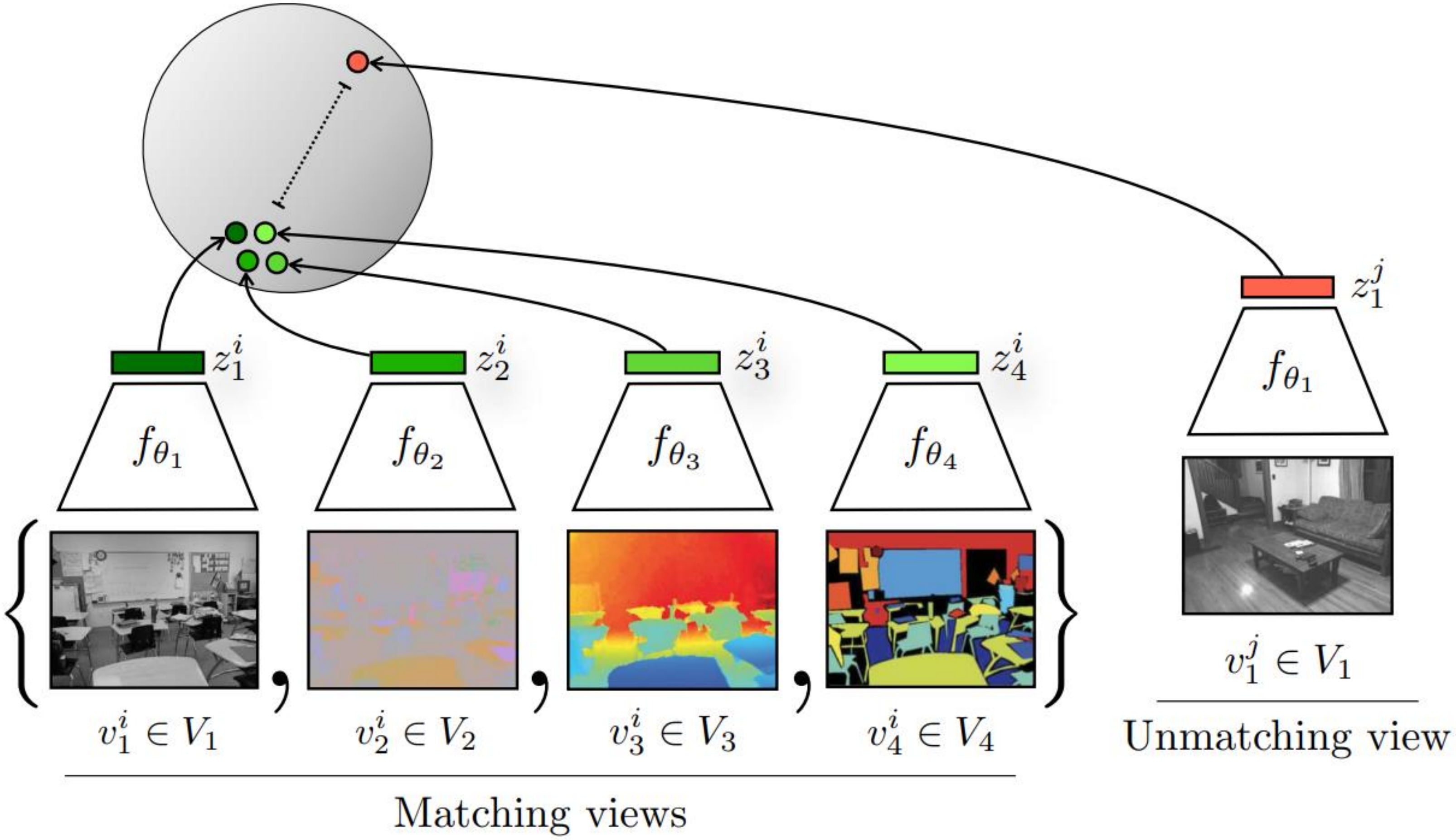
Contrast based: Context-Context Contrast

- Cluster-based Discrimination: DeepCluster [32], Local Aggregation [33]



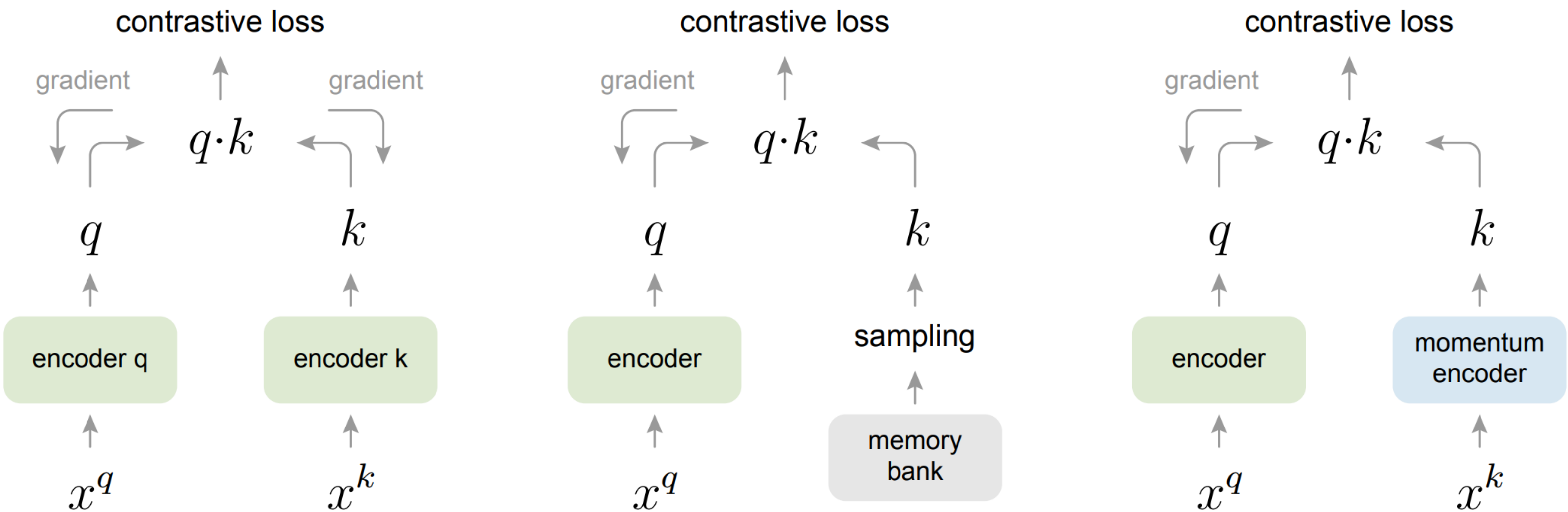
Contrast based: Context-Context Contrast

- Instance Discrimination: Contrastive Multiview Coding (CMC) [31]



Contrast based: Context-Context Contrast

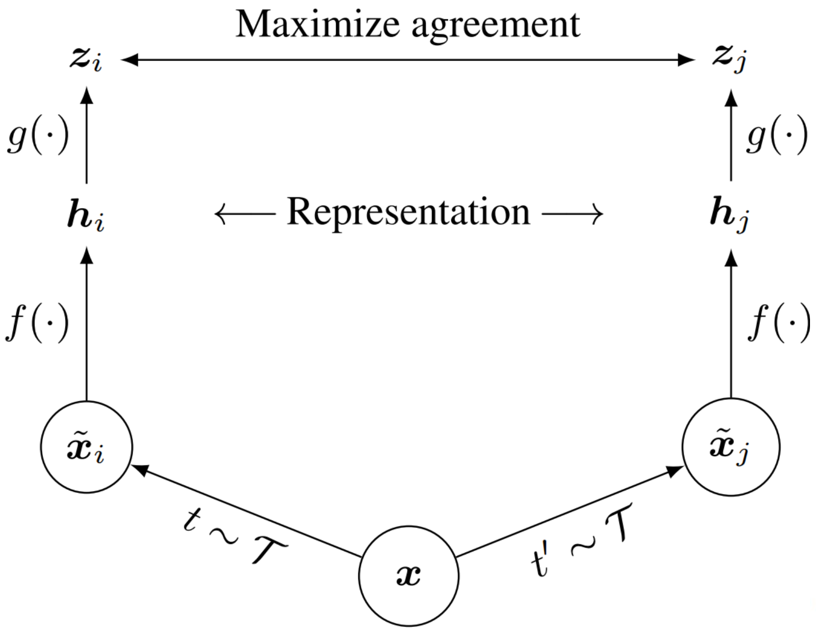
- Instance Discrimination: Momentum Contrast for Unsupervised Visual Representation Learning (MoCo) [34,37]

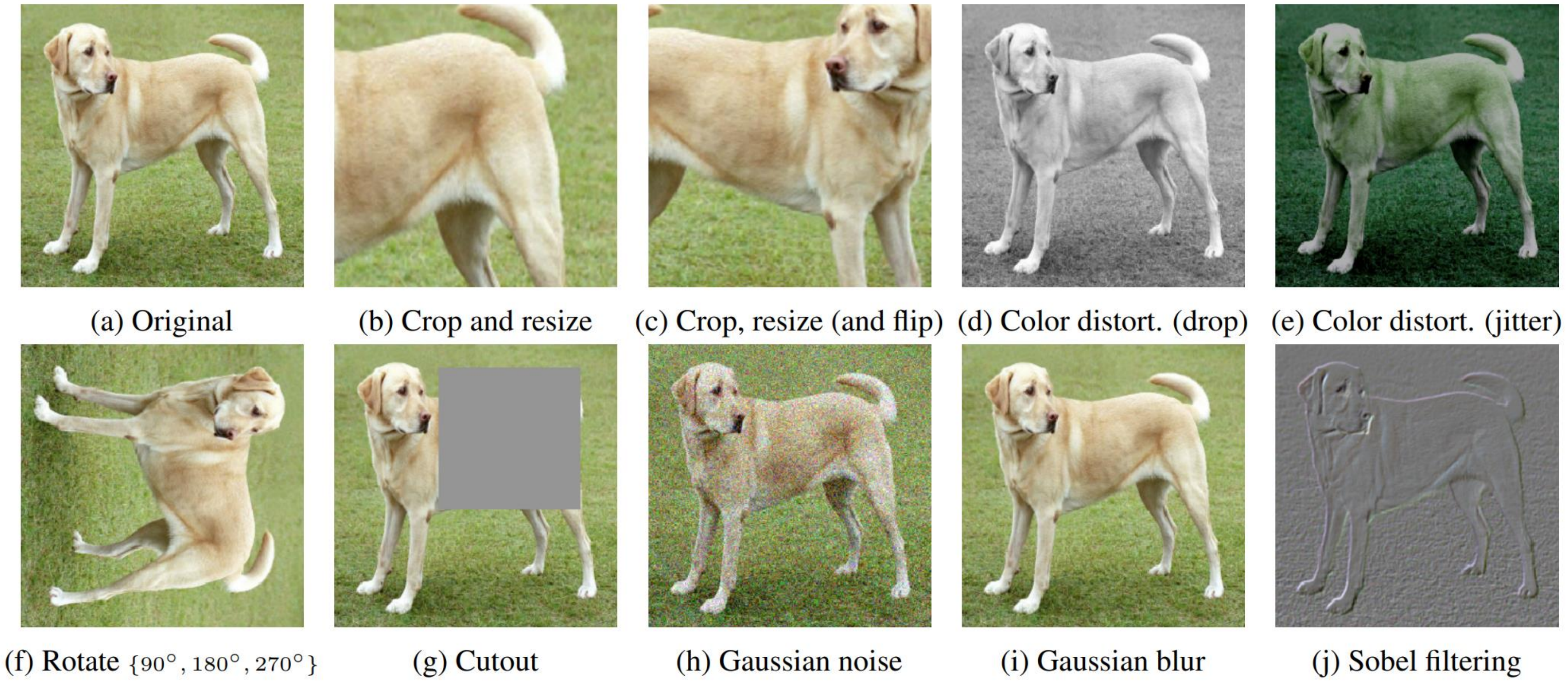


(a) end-to-end (b) memory bank (c) MoCo

Contrast based: Context-Context Contrast

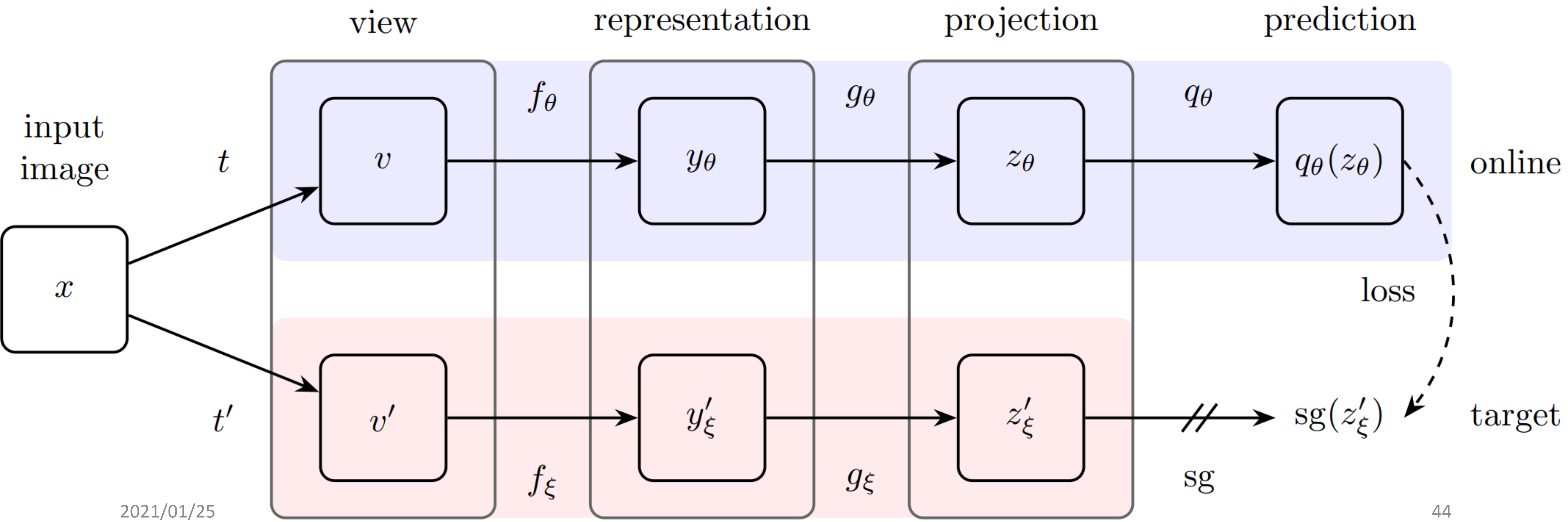
- Instance Discrimination: A Simple Framework for Contrastive Learning of Visual Representations (SimCLR) [35,38]





Contrast based: Context-Context Contrast

- Instance Discrimination: Bootstrap Your Own Latent A New Approach to Self-Supervised Learning (BYOL) [36]



Context based vs Contrast based

- Context based
  - Autoregressive Language Model (LM): ELMo [3], GPT-1/2/3 [4,5,6]
  - Denoising Auto-Encoder (DAE): MLM (BERT[7], RoBERTa[9], ERNIE[21,23], UniLM[14], XLM [15]), Seq2SeqMLM (MASS [11], T5 [17], ProphetNet [43], BART[12])
  - Permuted Language Model (PLM): XLNet [10], MPNet [27]
- Contrast based
  - Context-Instance Contrast
    - Predict Relative Position (PRP): Jigsaw, Rotation Angle [45], Sentence Order Prediction (ALBERT [19], StructBERT [20])
    - Maximize Mutual Information (MI): Deep InfoMax/InforWord [28], AMDIM [29], Contrastive Predictive Coding [30] (wav2vec [41,42]), Replaced Token Detection (word2vec [1], ELECTRA[18])
  - Context-Context Contrast
    - DeepCluster [32], CMC [31], MoCo [34,37], SimCLR [35,38], BYOL [36], Next Sentence Prediction (BERT [7])

How to use pre-training for downstream tasks?

- Choose pre-training task, model structure, data in pre-training
- In fine-tuning
  - Feature incorporation or fine-tunning
  - What to fine-tune? Embedding, partial layers, whole model
  - Different fine-tuning stages, layer-wise fine-tuning
  - Extra fine-tuning adaptors
- Reduce the gap between pre-training and fine-tuning
  - Different pre-training tasks for different downstream tasks
  - Make the data and model consistency with downstream tasks
  - Joint pre-training and fine-tuning

How to use pre-training for downstream tasks?

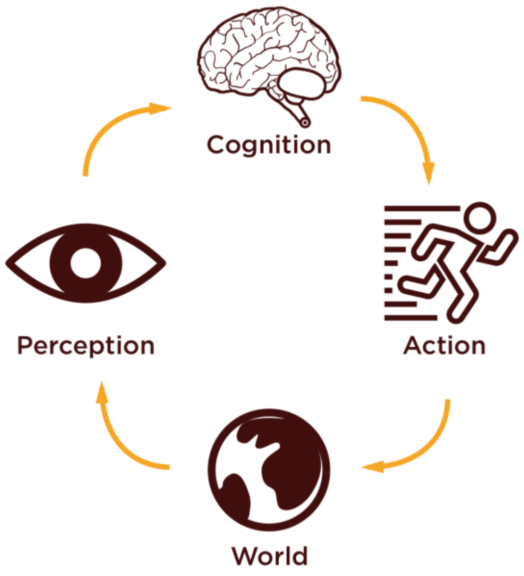
- Different pre-training tasks for different downstream tasks
  - CoLA: The Corpus of Linguistic Acceptability, prefer ELECTRA
  - RTE, MNLI, QNLI: prefer sentence pair pre-training, such as SOP
  - NER: prefer non-degeneration output hidden, BERT instead of ELECTRA
  - SQuAD: prefer span based prediction, span mask
  - NMT, text summarization: prefer seq2seqMLM or conditional sequence generation

How to use pre-training for downstream tasks?

- Compress the pre-trained model for practical deployment
  - Pruning: Compressing BERT [47], LayerDrop [48]
  - Quantization: Q-BERT [49], Q8BERT [50]
  - Parameter sharing: ALBERT[19]
  - Knowledge distillation: DistilBERT [51], TinyBERT [52], LightPAFF [53], BERT-PKD [54], MobileBERT [55], MiniLM [56], DynaBERT [57]
  - Neural architecture search: AdaBERT [58], NAS-BERT [59]

Comparison between pre-trained models for NLP, CV and Speech

- Model size and data size
  - Image: SimCLRv2/800M/300M images, DALL-E/12B/250M image-text pairs,
  - Speech: (Conformer + Wav2vec 2.0)/1B/60K hours speech data
  - NLP: GPT-3/175B/400B tokens → Switch Transformers/1.6 Trillion/180B tokens
- Context-based or Contrast-based?
  - Image, speech, more contrast based
  - NLP, more context-based
- Perception vs Cognition
  - Image, speech is more like perception
  - NLP is more like cognition



Summary of this course

- Overview of pre-training in NLP, CV and Speech
- Taxonomy of self-supervised based pre-training
    - Context based
    - Contrast based
- More discussion about pre-training
    - How to use for down-streaming tasks
    - Comparison between NLP, CV and Speech

Thank You!

Xu Tan
Senior Researcher @ Microsoft Research Asia
xuta@microsoft.com

https://www.microsoft.com/en-us/research/people/xuta/

Reference

[1] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In NeurIPS, 2013.

[2] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In EMNLP, 2014

[3] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In NAACL-HLT, 2018.

[4] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018. URL https://s3-us-west-2.amazonaws. com/openai-assets/researchcovers/languageunsupervised/ languageunderstandingpaper.pdf.

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 2019.

[6] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, 2019.

[8] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In NeurIPS, 2017.

[9] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

[10] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In NeurIPS, pages 5754–5764, 2019.

Reference

[11] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MASS: masked sequence to sequence pre-training for language generation. In ICML, volume 97 of Proceedings of Machine Learning Research, pages 5926–5936, 2019.

[12] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-tosequence pre-training for natural language generation, translation, and comprehension. arXiv preprint arXiv:1910.13461, 2019.

[13] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. arXiv preprint arXiv:2001.08210, 2020.

[14] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In NeurIPS, pages 13042–13054, 2019.

[15] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. In NeurIPS, pages 7057–7067, 2019.

[16] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard ´ Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116, 2019.

[17] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683, 2019.

[18] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR, 2020.

[19] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In International Conference on Learning Representations, 2020.

[20] Wei Wang, Bin Bi, Ming Yan, Chen Wu, Zuyi Bao, Liwei Peng, and Luo Si. StructBERT: Incorporating language structures into pre-training for deep language understanding. In ICLR, 2020.

Reference

[21] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. In ACL, 2019.

[22] Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. Unicoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In EMNLP-IJCNLP, pages 2485–2494, 2019

[23] Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. ERNIE: enhanced representation through knowledge integration. arXiv preprint arXiv:1904.09223, 2019.

[24] Yung-Sung Chuang, Chi-Liang Liu, and Hung-yi Lee. SpeechBERT: Cross-modal pre-trained language model for end-to-end spoken question answering. arXiv preprint arXiv:1910.11559, 2019.

[25] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

[26] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. TinyBERT: Distilling BERT for natural language understanding. arXiv preprint arXiv:1909.10351, 2019.

[27] Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. Y. Mpnet: Masked and permuted pre-training for language understanding. NeurIPS 2020.

[28] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio. Learning deep representations by mutual information estimation and maximization. arXiv preprint arXiv:1808.06670, 2018.

[29] P. Bachman, R. D. Hjelm, and W. Buchwalter. Learning representations by maximizing mutual information across views. In NIPS, pages 15509–15519, 2019.

[30] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748, 2018.

Reference

[31] Y. Tian, D. Krishnan, and P. Isola. Contrastive multiview coding. arXiv preprint arXiv:1906.05849, 2019.

[32] M. Caron, P. Bojanowski, A. Joulin, and M. Douze. Deep clustering for unsupervised learning of visual features. In Proceedings of the ECCV (ECCV), pages 132–149, 2018.

[33] C. Zhuang, A. L. Zhai, and D. Yamins. Local aggregation for unsupervised learning of visual embeddings. In Proceedings of the IEEE ICCV, pages 6002–6012, 2019.

[34] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. arXiv preprint arXiv:1911.05722, 2019.

[35] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020.

[36] J.-B. Grill, F. Strub, F. Altche, C. Tallec, P. H. Richemond, ´ E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. arXiv preprint arXiv:2006.07733, 2020.

[37] X. Chen, H. Fan, R. Girshick, and K. He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020

[38] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. Hinton. Big self-supervised models are strong semi-supervised learners. arXiv preprint arXiv:2006.10029, 2020.

[39] Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C. C., Pang, R., ... & Wu, Y. (2020). Pushing the Limits of Semi-Supervised Learning for Automatic Speech Recognition. *arXiv preprint arXiv:2010.10504*.

[40] Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., ... & Auli, M. (2020). Self-training and Pre-training are Complementary for Speech Recognition. *arXiv preprint arXiv:2010.11430*.

Reference

[41] Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised Pre-Training for Speech Recognition. *Proc. Interspeech 2019*, 3465-3469.

[42] Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

[43] Qi, W., Yan, Y., Gong, Y., Liu, D., Duan, N., Chen, J., ... & Zhou, M. (2020, November). ProphetNet: Predicting Future N-gram for Sequence-to-Sequence Pre-training. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings (pp. 2401-2410).

[44] Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 1-26.

[45] Liu, X., Zhang, F., Hou, Z., Wang, Z., Mian, L., Zhang, J., & Tang, J. (2020). Self-supervised learning: Generative or contrastive. arXiv preprint arXiv:2006.08218, 1(2).

[46] Fedus, W., Zoph, B., & Shazeer, N. (2021). Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv preprint arXiv:2101.03961.

[47] Mitchell A Gordon, Kevin Duh, and Nicholas Andrews. Compressing bert: Studying the effects of weight pruning on transfer learning. arXiv preprint arXiv:2002.08307, 2020.

[48] Angela Fan, Edouard Grave, and Armand Joulin. Reducing transformer depth on demand with structured dropout. In International Conference on Learning Representations, 2019.

[]49 Sheng Shen, Zhen Dong, Jiayu Ye, Linjian Ma, Zhewei Yao, Amir Gholami, Michael W Mahoney, and Kurt Keutzer. Q-bert: Hessian based ultra low precision quantization of bert. In AAAI, pp. 8815–8821, 2020.

[50] Ofir Zafrir, Guy Boudoukh, Peter Izsak, and Moshe Wasserblat. Q8bert: Quantized 8bit bert. arXiv preprint arXiv:1910.06188, 2019.

Reference

[51] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108, 2019.

[52] Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351, 2019.

[53] Kaitao Song, Hao Sun, Xu Tan, Tao Qin, Jianfeng Lu, Hongzhi Liu, and Tie-Yan Liu. Lightpaff: A two-stage distillation framework for pre-training and fine-tuning. arXiv preprint arXiv:2004.12817, 2020.

[54] Siqi Sun, Yu Cheng, Zhe Gan, and Jingjing Liu. Patient knowledge distillation for bert model compression. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP), pp. 4314–4323, 2019.

[55] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In Proceedings of the IEEE International Conference on Computer Vision, pp. 1314–1324, 2019.

[56] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep selfattention distillation for task-agnostic compression of pre-trained transformers. arXiv preprint arXiv:2002.10957, 2020b.

[57] Lu Hou, Lifeng Shang, Xin Jiang, and Qun Liu. Dynabert: Dynamic bert with adaptive width and depth. arXiv preprint arXiv:2004.04037, 2020.

[58] Daoyuan Chen, Yaliang Li, Minghui Qiu, Zhen Wang, Bofang Li, Bolin Ding, Hongbo Deng, Jun Huang, Wei Lin, and Jingren Zhou. Adabert: Task-adaptive bert compression with differentiable neural architecture search. In Christian Bessiere (ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, pp. 2463–2469.

[59] Jin Xu, Xu Tan, Renqian Luo, Kaitao Song, Jian Li, Tao Qin and Tie-Yan Liu, Task-Agnostic and Adaptive-Size BERT Compression, openreview 2021.